

ĐẠI HỌC ĐÀ NẴNG  
TRƯỜNG ĐẠI HỌC BÁCH KHOA  
KHOA CÔNG NGHỆ THÔNG TIN

**ĐỒ ÁN TỐT NGHIỆP**  
**NGÀNH: CÔNG NGHỆ THÔNG TIN**  
**CHUYÊN NGÀNH: KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ  
NHÂN TẠO**

ĐỀ TÀI:

**HỆ THỐNG HỎI ĐÁP THỰC VẬT RỪNG  
KHU VỰC ĐÀ NẴNG**

Người hướng dẫn: TS. NGUYỄN VĂN HIỆU  
Sinh viên thực hiện: HỒ QUỐC THIÊN ANH  
Số thẻ sinh viên: 102210089  
Lớp: 21TCLC\_KHDL

Đà Nẵng, 05/2025

**ĐẠI HỌC ĐÀ NẴNG  
TRƯỜNG ĐẠI HỌC BÁCH KHOA  
KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỒ ÁN TỐT NGHIỆP  
NGÀNH: CÔNG NGHỆ THÔNG TIN  
CHUYÊN NGÀNH: KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ  
NHÂN TẠO**

**ĐỀ TÀI:**

**HỆ THỐNG HỎI ĐÁP THỰC VẬT RỪNG  
KHU VỰC ĐÀ NẴNG**

Người hướng dẫn: TS. NGUYỄN VĂN HIỆU

Sinh viên thực hiện: HỒ QUỐC THIÊN ANH

Số thẻ sinh viên: 102210089

Lớp: 21TCLC\_KHDL

Đà Nẵng, 05/2025

## NHẬN XÉT CỦA NGƯỜI HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Dà Nẵng, ..... , 2025

Người hướng dẫn

## NHẬN XÉT CỦA NGƯỜI PHẢN BIỆN

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

Dà Nẵng, ..... , ..... , 2025

Người phản biện

## TÓM TẮT

Tên đề tài: Hệ thống Hỏi Đáp Thực vật Rừng Khu vực Đà Nẵng

Sinh viên thực hiện: HỒ QUỐC THIÊN ANH

Số thẻ SV: 102210089

Lớp: 21TCLC\_KHDL

Hệ thực vật rừng đóng vai trò thiết yếu trong đa dạng sinh học và cân bằng hệ sinh thái. Tuy nhiên, việc tiếp cận thông tin chính xác về các loài thực vật, đặc biệt tại các khu vực đặc thù như Đà Nẵng, còn nhiều thách thức do sự phân mảnh dữ liệu và rào cản chuyên môn.

Luận văn này giới thiệu hệ thống Hỏi đáp Thực vật Rừng Đà Nẵng - một nền tảng đa phương thức tích hợp thị giác máy tính (CV) và xử lý ngôn ngữ tự nhiên (NLP) nhằm hỗ trợ nhận diện và tra cứu thông tin thực vật một cách tự động. Đóng góp nổi bật của hệ thống là kiến trúc Dual-Stream Fusion tiên tiến, kết hợp mô hình học sâu cho phân loại hình ảnh với mô hình sinh ngôn ngữ tăng cường truy xuất (Retrieval-Augmented Generation, RAG) để trả lời câu hỏi ngữ cảnh bằng tiếng Việt.

Hệ thống khai thác cơ sở tri thức thực vật chuyên biệt, được xây dựng từ các nguồn khoa học đáng tin cậy. Khả năng của hệ thống được đánh giá trên bộ dữ liệu thực vật rừng khu vực Đà Nẵng, cho kết quả quan cả về độ chính xác phân loại lẫn mức độ hài lòng của người dùng.

Với giao diện web thân thiện và hiệu năng tốt, hệ thống có tiềm năng ứng dụng trong giáo dục, nghiên cứu và nâng cao nhận thức cộng đồng về bảo tồn đa dạng sinh học. Đây là bước tiến quan trọng trong việc ứng dụng AI vào lĩnh vực bảo tồn thực vật tại Việt Nam.

### NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Họ tên sinh viên: **Hồ Quốc Thiên Anh**

Số thẻ sinh viên: **102210089**

Lớp: **21TCLC\_KHDL**

Khoa: **Công nghệ Thông tin**

Ngành: **Khoa học dữ liệu và Trí tuệ nhân tạo**

1. *Tên đề tài đồ án: Hệ thống Hỏi Đáp Thực vật Rừng Khu vực Đà Nẵng.*
2. *Đề tài thuộc diện:  Có ký kết thỏa thuận sở hữu trí tuệ đối với kết quả thực hiện.*
3. *Các số liệu và dữ liệu ban đầu: Không có.*
4. *Nội dung các phần thuyết minh và tính toán:*
  - **Mở đầu:** Giới thiệu tổng quan về đề tài, bao gồm mục tiêu, ý nghĩa, phương pháp thực hiện, kết quả mong đợi và cấu trúc của đồ án.
  - **Chương 1. Tổng quan bài toán:** Trình bày tổng quan về vấn đề mà đồ án giải quyết, các khó khăn khi giải quyết vấn đề.
  - **Chương 2. Cơ sở lý thuyết:** Trình bày các kiến thức lý thuyết đã thu nhận và áp dụng trong đồ án.
  - **Chương 3. Giải pháp đề xuất:** Trình bày thiết kế của hệ thống, quá trình giải quyết vấn đề dựa trên lý thuyết đã đề cập ở chương 2.
  - **Chương 4. Triển khai và đánh giá kết quả:** Trình bày cách cài đặt và đánh giá kết quả đạt được.
5. *Các bản vẽ, đồ thị (ghi rõ các loại và kích thước bản vẽ): Không có.*
6. *Họ tên người hướng dẫn: TS. Nguyễn Văn Hiệu*
7. *Ngày giao nhiệm vụ đồ án: ...../...../2025*
8. *Ngày hoàn thành đồ án: ...../...../2025*

Đà Nẵng, ..., ..., 2025

Trưởng Bộ môn .....

Người hướng dẫn

## LỜI NÓI ĐẦU

Đề tài “Hệ thống Hỏi Đáp Thực vật Rừng Khu vực Đà Nẵng” là đề tài mà em đặt nhiều tâm huyết và nỗ lực của bản thân để hoàn thiện và đây cũng là đề tài đồ án tốt nghiệp của bản thân để hoàn thành chương trình học tại Khoa Công nghệ Thông tin - Trường Đại học Bách khoa – Đại học Đà Nẵng.

Em xin chân thành gửi lời cảm ơn đến Nhà trường và quý thầy cô trong Khoa Công nghệ Thông tin đã truyền đạt những kiến thức và kỹ năng quý giá trong quá trình học tập tại đây, là nền tảng vững chắc giúp em thực hiện và hoàn thiện đồ án này.

Đặc biệt, em xin gửi lời cảm ơn sâu sắc nhất đến thầy Nguyễn Văn Hiệu, người đã dành thời gian và tận tình hướng dẫn, hỗ trợ em rất nhiều trong suốt quá trình thực hiện để em có thể hoàn thành đồ án này đúng tiến độ. Sự hỗ trợ, chia sẻ kinh nghiệm và sự hướng dẫn của thầy là sự đóng góp quan trọng, giúp em có thể áp dụng kiến thức và hoàn thành đề tài một cách thành công.

Quá trình thực hiện và nội dung trình bày ở đây chắc chắn sẽ không tránh khỏi những thiếu sót nhất định. Em rất mong sẽ nhận được sự thông cảm, góp ý và tận tình chỉ bảo của quý thầy cô, quý anh chị và các bạn để đề tài được hoàn thiện hơn.

Sau cùng, em xin kính chúc quý thầy cô, quý anh chị và các bạn dồi dào sức khỏe, chúc Khoa Công nghệ Thông tin và quý nhà trường ngày càng phát triển mạnh mẽ.

## CAM ĐOAN

Tôi xin cam đoan đồ án tốt nghiệp này là kết quả của quá trình nghiên cứu và làm việc độc lập của tôi, dưới sự hướng dẫn và giám sát của TS. Nguyễn Văn Hiệu. Tôi xin cam kết những điều sau:

- Tính độc đáo và Trích dẫn:** Mọi thông tin, dữ liệu, phân tích và kết luận trình bày trong luận văn này là nguyên bản và được thu thập thông qua các phương pháp nghiên cứu hợp lệ. Mọi nguồn tài liệu bên ngoài được sử dụng đều đã được trích dẫn và tham chiếu một cách thích hợp theo tiêu chuẩn học thuật.
- Không đạo văn:** Luận văn này không chứa nội dung sao chép nguyên văn từ các nguồn khác mà không ghi rõ nguồn. Mọi ý tưởng và đóng góp từ các nguồn bên ngoài đều đã được ghi nhận và tham chiếu.
- Làm việc độc lập:** Mặc dù nhận được sự hướng dẫn và phản hồi quý báu từ TS. Nguyễn Văn Hiệu, tôi đã tự thực hiện quá trình nghiên cứu, thiết kế, triển khai và viết luận văn này một cách độc lập.
- Trách nhiệm về tính chính xác:** Tôi hoàn toàn chịu trách nhiệm về tính chính xác và toàn vẹn của thông tin được trình bày trong luận văn này. Tôi sẽ chịu trách nhiệm cho bất kỳ hành vi thiếu trung thực nào trong học thuật, bao gồm đạo văn hoặc vi phạm bản quyền, nếu bị phát hiện trong công trình này. Tôi hoàn toàn chấp nhận trách nhiệm về tính toàn vẹn của lời cam đoan này.

Đà Nẵng, ..... , 2025

Sinh viên thực hiện

Hồ Quốc Thiên Anh

## MỤC LỤC

<b>TÓM TẮT .....</b>	iii
<b>NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP.....</b>	iv
<b>LỜI NÓI ĐẦU.....</b>	i
<b>CAM ĐOAN.....</b>	ii
<b>MỤC LỤC .....</b>	iii
<b>DANH SÁCH BẢNG.....</b>	vi
<b>DANH SÁCH HÌNH VẼ .....</b>	vii
<b>DANH SÁCH VIẾT TẮT .....</b>	viii
<b>MỞ ĐẦU.....</b>	1
<b>CHƯƠNG 1: TỔNG QUAN BÀI TOÁN .....</b>	2
<b>1.1 Đa dạng sinh học thực vật và thách thức tiếp cận thông tin.....</b>	2
1.1.1 Tâm quan trọng của đa dạng sinh học thực vật rừng .....	2
1.1.2 Thách thức hiện tại trong nhận diện và phân loại thực vật .....	2
<b>1.2 Hệ thống hỏi đáp thực vật: Định nghĩa và phạm vi.....</b>	4
1.2.1 Khái niệm hệ thống hỏi đáp thực vật .....	4
1.2.2 Thách thức đặc thù trong lĩnh vực thực vật học .....	5
<b>1.3. Các nghiên cứu liên quan và hiện trạng.....</b>	7
1.2.3 Hệ thống nhận diện thực vật hiện có .....	7
1.2.4 Hệ thống hỏi đáp chuyên biệt cho thực vật .....	8
1.2.5 Ứng dụng AI trong bảo tồn sinh học .....	8
1.2.6 Mục tiêu nghiên cứu .....	9
<b>CHƯƠNG 2. CƠ SỞ LÝ THUYẾT .....</b>	10
<b>2.1. Kiến trúc Deep Learning cho nhận diện thực vật .....</b>	10
2.1.1. ConvNeXt V2 .....	10

2.1.2. Vision Transformers .....	14
2.1.3. Dual-Stream Fusion Architecture .....	19
<b>2.2. Metric learning và ProxyNCA loss .....</b>	<b>22</b>
2.2.1 Lý thuyết phân loại dựa trên embedding.....	22
2.2.2. ProxyNCA loss .....	23
<b>2.3. Tìm kiếm tương đồng và phát hiện mẫu ngoài phân phối (OOD).....</b>	<b>26</b>
2.3.1. FAISS và bài toán tìm kiếm tương đồng trong không gian nhiều chiều .....	26
2.3.2. Phát hiện mẫu ngoài phân phối (OOD) bằng phương pháp trung bình khoảng cách theo lớp.....	28
<b>2.4. Xử lý ngôn ngữ Việt Nam cho hệ thống hỏi đáp về thực vật.....</b>	<b>33</b>
2.4.1. Lý thuyết ngữ nghĩa phân bố (Distributional Semantics) .....	33
2.4.2. Không gian vector và tương tự ngữ nghĩa.....	33
2.4.3. Retrieval-Augmented Generation.....	34
2.4.4. Lý thuyết xử lý đa ngôn ngữ .....	36
2.4.5. Những thách thức lý thuyết đặc thù với tiếng Việt.....	37
2.4.6. Mô hình ngôn ngữ lớn (Large Language Models) .....	37
<b>CHƯƠNG 3: GIẢI PHÁP ĐỀ XUẤT.....</b>	<b>39</b>
<b>3.1. Tổng quan kiến trúc của hệ thống .....</b>	<b>39</b>
<b>3.2. Truy xuất ảnh thực vật.....</b>	<b>41</b>
<b>3.3. Phân loại thực vật .....</b>	<b>42</b>
<b>3.4. Truy xuất văn bản.....</b>	<b>43</b>
<b>3.5. Sinh văn bản.....</b>	<b>48</b>
<b>CHƯƠNG 4: TRIỂN KHAI VÀ ĐÁNH GIÁ KẾT QUẢ.....</b>	<b>51</b>
<b>4.1. Thực nghiệm huấn luyện.....</b>	<b>51</b>
4.1.1. Bộ dữ liệu .....	51
4.1.2 Cấu hình huấn luyện.....	52
<b>4.2. Đánh giá kết quả .....</b>	<b>53</b>

4.2.1. Đánh giá kết quả mô hình phân loại.....	53
4.2.2 Đánh giá kết quả mô hình truy xuất ảnh .....	54
4.2.3. Đánh giá kết quả mô hình truy xuất văn bản.....	56
<b>4.3. Triển khai .....</b>	<b>57</b>
4.3.1. Giao diện ứng dụng .....	58
4.3.2. Hỏi đáp .....	58
4.3.3. Thư viện.....	61
<b>KẾT LUẬN .....</b>	<b>62</b>
<b>TÀI LIỆU THAM KHẢO.....</b>	<b>65</b>

## **DANH SÁCH BẢNG**

Bảng 4-1: Cấu hình huấn luyện các mô hình xử lý ảnh .....	52
Bảng 4-2: Bảng đánh giá kết quả mô hình phân loại .....	53
Bảng 4-3: Bảng đánh giá kết quả mô hình truy xuất ảnh .....	55
Bảng 4-5: Bảng đánh giá kết quả mô hình truy xuất văn bản .....	57

## DANH SÁCH HÌNH VẼ

Hình 2-1: Thuật toán Global Response Normalization .....	11
Hình 2-2: So sánh kiến trúc của ResNet và ConvNextV2 .....	12
Hình 2-3: Kiến trúc Vision Transformer [31].....	15
Hình 2-4: Minh họa attention của các head trong ViT với cây Tai mèo .....	16
Hình 2-5: Minh họa attention của các head trong ViT với cây Rau má tía .....	17
Hình 2-6: Kiến trúc khối Dual-Stream Fusion .....	20
Hình 2-7: Mô tả quá trình cập nhật của Proxy NCA loss trong bài toán thực vật học.....	25
Hình 2-8: Minh họa thuật toán phát hiện OOD.....	31
Hình 3-1: Sơ đồ tổng quan luồng hoạt động của hệ thống .....	40
Hình 3-2: Luồng xử lý câu hỏi của FloraQA .....	50
Hình 4-1: Biểu đồ phân phối số lượng ảnh theo loài - họ - ngành trong tập dữ liệu .....	51
Hình 4-2: Feature map từ các mô hình khác nhau.....	53
Hình 4-3: Đường cong ROC của mô hình truy xuất ảnh với backbone Dual-Stream Fusion .....	55
Hình 4-4: Minh họa phân bố của 20 lớp thực vật ngẫu nhiên tạo bởi mô hình truy xuất ảnh .....	56
Hình 4-5: Trang chủ.....	58
Hình 4-6 Giao diện tính năng hỏi đáp .....	59
Hình 4-7: Giao diện tính năng hỏi đáp - 2 .....	60
Hình 4-8: Giao diện tính năng thư viện.....	61

## DANH SÁCH VIẾT TẮT

Từ viết tắt	Điễn giải
AI	Artificial Intelligence
CV	Computer Vision
NLP	Natural Language Processing
CNN	Convolution Neural Network
RNN	Recurrent Neural Network
ViT	Vision Transformers
FAISS	Facebook AI Similarity Search
QA	Question Answering
GRN	Global Response Normalization
MAE	Masked Autoencoder
OOD	Out of Distribution
LLM	Large Language Model
RRF	Reciprocal Rank Fusion
RAG	Retrieval Augmented Generation

## MỞ ĐẦU

Việc nhận dạng thực vật và tiếp cận kiến thức thực vật học là những rào cản quan trọng đối với công tác bảo tồn đa dạng sinh học. Sự đa dạng của các loài thực vật, sự biến đổi của hình dạng và sự phức tạp của thông tin thực vật học đặt ra những thách thức đáng kể không chỉ cho các nhà nghiên cứu mà còn cho công chúng muốn tìm hiểu về thực vật rừng.

Trong công trình này, tôi trình bày một hệ thống hỏi-đáp thực vật được thiết kế riêng cho khu vực rừng Đà Nẵng. Hệ thống này tận dụng sức mạnh của các mô hình học sâu, kết hợp các kỹ thuật thị giác máy tính để nhận dạng chính xác loài thực vật với xử lý ngôn ngữ tự nhiên để truy xuất thông tin. Nó cũng sử dụng kiến trúc hợp nhất luồng kép, kết hợp các mô hình học sâu để phân loại tốt hơn, tích hợp với kỹ thuật sinh tăng cường truy xuất (retrieval-augmented generation) để trả lời câu hỏi theo ngữ cảnh bằng tiếng Việt.

Một giao diện web thân thiện với người dùng cùng cơ sở dữ liệu thực vật chi tiết đã được phát triển để hỗ trợ bảo tồn rừng tại khu vực Đà Nẵng.

Luận văn này được cấu trúc như sau:

**Chương 1: Tổng quan bài toán** định nghĩa vấn đề xây dựng một hệ thống hỏi-đáp thực vật, những thách thức và mục tiêu nghiên cứu.

**Chương 2: Cơ sở lý thuyết** trình bày các nền tảng lý thuyết xây dựng hệ thống, bao gồm các mô hình học sâu cho các mô-đun: phân loại thực vật, truy xuất ảnh, truy xuất văn bản và sinh văn bản cũng các phương pháp đánh giá.

**Chương 3: Giải pháp đề xuất** chi tiết hóa luồng hệ thống và các thuật toán trong mô-đun phân loại thực vật, hệ thống hỏi-đáp và các thành phần tích hợp cơ sở dữ liệu.

**Chương 4: Triển khai và đánh giá kết quả** trình bày thiết lập thực nghiệm và các kết quả thu được từ việc triển khai hệ thống và so sánh với các phương pháp khác.

**Kết luận** tóm tắt những đóng góp chính, các hạn chế và hướng phát triển tiềm năng của nghiên cứu này.

## CHƯƠNG 1: TỔNG QUAN BÀI TOÁN

### 1.1 Đa dạng sinh học thực vật và thách thức tiếp cận thông tin

#### 1.1.1 *Tầm quan trọng của đa dạng sinh học thực vật rừng*

Hệ sinh thái rừng đóng vai trò then chốt trong việc duy trì đa dạng sinh học toàn cầu và cung cấp các dịch vụ hệ sinh thái thiết yếu cho sự phồn vinh của loài người. Rừng che phủ gần một phần ba diện tích đất liền trên Trái Đất và chứa đựng hơn 80% đa dạng sinh học trên cạn [1]. Những hệ sinh thái này hoạt động như các bể chứa carbon tự nhiên, hấp thụ khoảng 2,6 tỷ tấn carbon dioxide mỗi năm, đóng vai trò quan trọng trong việc giảm thiểu biến đổi khí hậu [2]. Tầm quan trọng của đa dạng sinh học rừng vượt xa việc lưu trữ carbon, bao gồm sản xuất sinh khối, cung cấp môi trường sống, thu phấn, phát tán hạt giống, điều hòa khí hậu và các dịch vụ văn hóa.

Việt Nam được xếp hạng thứ 16 về đa dạng sinh học trên thế giới, minh chứng cho sự phong phú đặc biệt của đa dạng sinh học rừng [3]. Đất nước này sở hữu khoảng 12.000 loài thực vật, phân bố trên các hệ sinh thái đa dạng từ rừng núi mát ở sườn nam dãy Himalaya đến rừng mưa nhiệt đới và rừng ngập mặn ven biển. Dãy núi Trường Sơn, tạo thành xương sống của Việt Nam, đã được công nhận như Amazon của châu Á do đa dạng sinh học đáng kinh ngạc [3]. Khu vực này hỗ trợ nhiều loài đặc hữu và đóng vai trò như kho tàng tài nguyên di truyền quan trọng.

Tầm quan trọng kinh tế và xã hội của đa dạng thực vật rừng không thể phủ nhận. Hơn 1,6 tỷ người trên thế giới phụ thuộc trực tiếp vào rừng để sinh sống, trong khi các sản phẩm từ rừng tạo thành một phần không thể thiếu trong cuộc sống hàng ngày [4]. Tại Việt Nam, hệ thống y học cổ truyền từ lâu đã dựa vào sự đa dạng phong phú của thực vật rừng, với các cộng đồng bản địa sở hữu kiến thức sâu rộng về đặc tính chữa bệnh của hệ thực vật địa phương. Kiến thức sinh thái truyền thống này, được tích lũy và truyền qua nhiều thế hệ, thể hiện sự hiểu biết tinh vi về tương tác thực vật-môi trường và các thực hành quản lý tài nguyên bền vững [5].

#### 1.1.2 *Thách thức hiện tại trong nhận diện và phân loại thực vật*

Mặc dù tầm quan trọng cốt yếu của đa dạng sinh học thực vật, vẫn tồn tại những thách thức đáng kể trong việc tiếp cận và sử dụng thông tin thực vật học một cách hiệu

quả. Các phương pháp nhận diện thực vật truyền thống vẫn phức tạp, tốn thời gian và thường gây khó khăn cho những người không chuyên do phụ thuộc vào thuật ngữ thực vật học chuyên biệt [6]. Sự phức tạp này tạo ra rào cản đáng kể cho những người mới bắt đầu muốn tiếp thu kiến thức về các loài, góp phần vào hiện tượng mà các nhà nghiên cứu gọi là "sự tuyệt chủng của giáo dục thực vật học" [7].

Những thách thức trong nhận diện thực vật càng trở nên phức tạp do tính chất phức tạp vốn có của hình thái học thực vật. Ngay cả các nhà thực vật học giàu kinh nghiệm cũng gặp khó khăn trong việc phân biệt các loài họ hàng gần có thể chỉ khác nhau ở những đặc điểm tinh vi [6]. Việc nhận diện thực vật trở nên đặc biệt khó khăn khi mẫu vật ở trạng thái sinh dường, vì các biến đổi hình thái do điều kiện môi trường, giai đoạn sinh học và đa dạng trong loài có thể gây nhầm lẫn trong nỗ lực nhận diện.

Rào cản ngôn ngữ và văn hóa càng làm phức tạp thêm việc tiếp cận thông tin thực vật học. Danh pháp khoa học, dù cung cấp sự chuẩn hóa, thường thiếu kết nối với hệ thống đặt tên địa phương và bối cảnh văn hóa. Sự ngắt kết nối này tạo ra khó khăn trong việc liên kết kiến thức khoa học với các thực hành truyền thống và nỗ lực bảo tồn địa phương. Hơn nữa, tính khả dụng hạn chế của các tài nguyên thực vật học bằng ngôn ngữ địa phương hạn chế khả năng tiếp cận thông tin thực vật học cho các cộng đồng có thể hưởng lợi nhiều nhất và đóng góp vào nỗ lực bảo tồn thực vật.

Việc phân mảnh kiến thức thực vật học đặt ra một thách thức đáng kể khác. Thông tin khoa học về các loài thực vật bị phân tán trên nhiều nguồn khác nhau, thường không thể tiếp cận được đối với những người thực hành cần nó nhất. Các hệ thống kiến thức truyền thống, vốn từ lâu đã cung cấp sự hiểu biết toàn diện về hệ thực vật địa phương, đang đối mặt với các mối đe dọa từ hiện đại hóa, đô thị hóa và việc di cư của những người nắm giữ kiến thức từ nông thôn ra thành thị [8]. Sự xói mòn kiến thức này đặc biệt đáng lo ngại vì kiến thức sinh thái truyền thống thường đại diện cho hàng thiên nhiên ký quan sát thực nghiệm và các thực hành quản lý thích ứng đã duy trì việc bảo tồn đa dạng sinh học.

### **1.1.3. Bối cảnh đặc thù của khu vực rừng Đà Nẵng**

Đà Nẵng đại diện cho một khu vực độc đáo và quan trọng cho việc bảo tồn đa dạng sinh học rừng ở miền Trung Việt Nam. Các khu bảo tồn thiên nhiên của thành phố, bao gồm rừng Sơn Trà, Bà Nà - Núi Chúa và Nam Hải Vân, đóng vai trò như "lá phổi xanh"

cho khu vực, cung cấp hàng triệu tấn oxy cho cư dân và du khách đồng thời chúa đựng đa dạng sinh học đặc biệt [9]. Riêng Khu bảo tồn thiên nhiên Sơn Trà đã ghi nhận 387 loài động vật và 1.010 loài thực vật, trong khi Khu bảo tồn Bà Nà - Núi Chúa có 626 loài động vật và 793 loài thực vật.

Vị trí địa lý đặc biệt của Đà Nẵng, nằm tại điểm giao thoa giữa các vùng sinh khí hậu khác nhau, tạo ra sự đa dạng cao về môi trường sống và loài. Khu vực này không chỉ quan trọng về mặt sinh thái mà còn có giá trị văn hóa sâu sắc, với nhiều loài thực vật được sử dụng trong y học cổ truyền và có ý nghĩa tinh thần đối với cộng đồng địa phương. Tuy nhiên, áp lực từ phát triển đô thị, du lịch và biến đổi khí hậu đang đặt ra những thách thức ngày càng tăng đối với việc bảo tồn đa dạng sinh học thực vật trong khu vực.

Sự phát triển nhanh chóng của Đà Nẵng như một trung tâm kinh tế và du lịch quan trọng đã tạo ra áp lực lớn lên các hệ sinh thái rừng tự nhiên. Việc mở rộng cơ sở hạ tầng, phát triển du lịch và các hoạt động kinh tế khác đòi hỏi cần có những chiến lược quản lý bền vững để cân bằng giữa phát triển kinh tế và bảo tồn đa dạng sinh học. Trong bối cảnh này, việc phát triển các công cụ và hệ thống hỗ trợ việc nhận diện, quản lý và bảo tồn đa dạng thực vật trở nên cấp thiết hơn bao giờ hết.

## **1.2 Hệ thống hỏi đáp thực vật: Định nghĩa và phạm vi**

### **1.2.1 Khái niệm hệ thống hỏi đáp thực vật**

Hệ thống hỏi đáp (Question Answering - QA) là một lĩnh vực con của khoa học máy tính trong xử lý ngôn ngữ tự nhiên và truy xuất thông tin, chuyên về việc phát triển các hệ thống có khả năng trả lời các câu hỏi được đặt ra bằng ngôn ngữ tự nhiên [10]. Các hệ thống này xác định ngữ cảnh đằng sau câu hỏi, trích xuất thông tin liên quan từ lượng lớn dữ liệu và trình bày lại cho người dùng một cách súc tích và dễ đọc. Trong bối cảnh thực vật học, hệ thống hỏi đáp thực vật được định nghĩa như một hệ thống chuyên biệt có khả năng hiểu và xử lý các truy vấn liên quan đến thực vật, từ nhận diện loài đến thông tin về đặc tính sinh học, phân bố địa lý và ứng dụng.

Hệ thống hỏi đáp thực vật hiện đại có thể được phân loại dựa trên các phương thức tương tác khác nhau mà chúng hỗ trợ. Các hệ thống dựa trên văn bản (text-based) cho phép người dùng đặt câu hỏi bằng ngôn ngữ tự nhiên về các đặc điểm thực vật, như "Loài cây nào có lá hình tim và hoa màu đỏ?". Loại hệ thống này dựa vào các kỹ thuật xử lý

ngôn ngữ tự nhiên để hiểu ý định của người dùng và truy xuất thông tin từ cơ sở dữ liệu kiến thức [11].

Các hệ thống hỏi đáp trực quan (Visual QA) được thiết kế để trả lời câu hỏi về hình ảnh, kết hợp kỹ thuật xử lý ngôn ngữ tự nhiên với thị giác máy tính [12]. Trong lĩnh vực thực vật học, loại hệ thống này cho phép người dùng tải lên hình ảnh của một loài thực vật và đặt câu hỏi như "Đây là loài cây gì?" hoặc "Cây này có thể ăn được không?". Sự phát triển của các mô hình đa phương thức (multimodal models) đã tạo ra khả năng xử lý đồng thời cả thông tin hình ảnh và văn bản, mang lại hiệu quả cao hơn trong nhận diện và phân loại thực vật [13].

Hệ thống hỏi đáp đa phương thức (hybrid/multimodal) đại diện cho bước tiến quan trọng nhất, cho phép tích hợp nhiều nguồn thông tin khác nhau để đưa ra câu trả lời chính xác hơn. Nghiên cứu cho thấy việc sử dụng thông tin đa nguồn chính xác hơn so với việc chỉ dựa vào một phương thức hình ảnh duy nhất, nhấn mạnh tính chất bổ sung của thông tin trực quan và mô tả văn bản trong quá trình nhận diện hình ảnh cây bệnh [14].

Hệ thống hỏi đáp thực vật khác biệt đáng kể so với các hệ thống hỏi đáp tổng quát về nhiều khía cạnh. Trong khi hệ thống hỏi đáp miền mở (open-domain QA) được thiết kế để xử lý câu hỏi về hầu hết mọi chủ đề và dựa vào kiến thức chung rộng lớn, hệ thống hỏi đáp thực vật thuộc loại hệ thống miền đóng (closed-domain QA) chuyên biệt trong lĩnh vực cụ thể [10]. Điều này cho phép chúng cung cấp câu trả lời chi tiết và chính xác hơn được điều chỉnh theo lĩnh vực thực vật học.

Một đặc điểm quan trọng khác là yêu cầu về độ chính xác cao trong nhận diện loài. Trong khi hệ thống QA tổng quát có thể chấp nhận mức độ mơ hồ nhất định trong câu trả lời, hệ thống QA thực vật đòi hỏi độ chính xác tuyệt đối trong việc nhận diện loài, vì sai lầm có thể dẫn đến hậu quả nghiêm trọng, đặc biệt trong các ứng dụng y học hoặc an toàn thực phẩm [15]. Hơn nữa, hệ thống QA thực vật thường cần xử lý các truy vấn phức tạp đòi hỏi kiến thức chuyên sâu về hình thái học, sinh lý học, sinh thái học và phân loại học.

### **1.2.2 Thách thức đặc thù trong lĩnh vực thực vật học**

Lĩnh vực thực vật học đặt ra những thách thức riêng biệt đối với các hệ thống hỏi đáp do sự đa dạng phong phú về hình thái của thực vật. Các loài thực vật có thể biến đổi đáng kể về hình dạng, kích thước, màu sắc và cấu trúc tùy thuộc vào giai đoạn phát triển, điều kiện môi trường và mùa trong năm [16]. Ví dụ, một loài cây có thể hiện hình thái

rất khác biệt giữa giai đoạn non và trưởng thành, hoặc giữa mùa ra hoa và các thời điểm khác.

Nhận diện thực vật càng trở nên phức tạp khi có sự tồn tại của các loài họ hàng gần với ngoại hình rất giống nhau. Ngay cả các nhà thực vật học giàu kinh nghiệm cũng có thể gặp khó khăn khi phân biệt các loài chỉ khác nhau ở những đặc điểm rất tinh vi, không dễ nhận diện qua hình ảnh thông thường [16]. Một số đặc trưng quan trọng, như cấu trúc chi tiết của hoa hoặc quả, có thể không luôn hiện diện trong ảnh chụp, làm tăng thêm độ khó của bài toán nhận diện.

Hệ thống hỏi đáp thực vật cần tích hợp kiến thức từ nhiều lĩnh vực khoa học khác nhau, điều này tạo ra thách thức đáng kể trong việc xây dựng một cơ sở tri thức toàn diện. Thông tin về một loài thực vật có thể liên quan đến dữ liệu từ các lĩnh vực như thực vật học (phân loại, hình thái), sinh thái học (môi trường sống, tương tác sinh học), sinh lý học (chức năng, trao đổi chất), di truyền học (DNA, tiến hóa), dược học (hoạt chất, độc tính), và nông học (canh tác, năng suất) [17].

Việc tích hợp hiệu quả các nguồn kiến thức đa dạng này đòi hỏi phải xử lý các vấn đề về tính nhất quán, độ tin cậy và khả năng cập nhật thông tin. Các cơ sở dữ liệu khác nhau có thể sử dụng các hệ thống phân loại, thuật ngữ và tiêu chuẩn dữ liệu khác nhau, dẫn đến khó khăn trong việc hợp nhất và diễn giải thông tin một cách chính xác.Thêm vào đó, kiến thức về thực vật học liên tục được cập nhật thông qua các nghiên cứu mới, đòi hỏi hệ thống phải có khả năng thích ứng và cập nhật linh hoạt.

Một trong những thách thức lớn nhất trong hệ thống hỏi đáp thực vật là xử lý sự đa dạng và phức tạp của hệ thống tên gọi thực vật. Danh pháp thực vật học (botanical nomenclature) là hệ thống đặt tên chính thức cho các loài thực vật, được quản lý theo Bộ quy tắc Danh pháp Quốc tế cho Tảo, Nấm và Thực vật (ICNafp) [18]. Mỗi loài thực vật có một tên khoa học duy nhất theo hệ thống phân loại nhị phân của Linnaeus, bao gồm tên chi (genus) và tên loài (specific epithet). Tuy nhiên, cùng một loài thực vật có thể được gọi bằng nhiều tên dân gian khác nhau, tùy theo khu vực địa lý, ngôn ngữ và văn hóa bản địa. Ví dụ, tên dân gian “cúc” có thể dùng để chỉ ít nhất 18 loài khác nhau [19]. Ngược lại, một tên dân gian có thể đồng thời chỉ nhiều loài thực vật khác nhau, dễ gây ra sự nhầm lẫn. Thách thức này càng tăng lên khi hệ thống phải xử lý các truy vấn đa ngôn ngữ và đồng thời liên kết các tên gọi khác nhau với cùng một thực thể thực vật.

Bên cạnh đó, danh pháp thực vật học có thể thay đổi theo thời gian do những tiến bộ trong nghiên cứu di truyền và phân loại học, dẫn đến việc tái phân loại và thay đổi tên gọi [18]. Điều này đòi hỏi hệ thống phải có khả năng xử lý các tên đồng nghĩa (synonyms), theo dõi và cập nhật các thay đổi trong phân loại học. Một thách thức khác là người dùng có thể nhập các tên gọi cũ, tên không chính thức hoặc thậm chí là tên bị viết sai chính tả, đòi hỏi hệ thống cần có cơ chế xử lý lỗi và hiểu ngữ cảnh một cách thông minh để cung cấp kết quả chính xác.

### **1.3. Các nghiên cứu liên quan và hiện trạng**

#### **1.2.3 Hệ thống nhận diện thực vật hiện có**

Trong lĩnh vực nhận diện thực vật tự động, PlantNet đã trở thành một trong những hệ thống tiên phong được phát triển bởi các viện nghiên cứu hàng đầu của Pháp, bao gồm IRD, CIRAD, INRA, INRIA và mạng lưới Tela Botanica [20]. Hệ thống này sử dụng phương pháp cộng tác khoa học công dân (citizen science) và machine learning để xây dựng cơ sở dữ liệu hình ảnh thực vật toàn cầu. Nghiên cứu đánh giá độc lập cho thấy PlantNet đạt độ chính xác khoảng 40% trong nhận diện chính xác và 67% khi tính cả các câu trả lời hữu ích một phần [21]. iNaturalist đại diện cho mô hình khác với hơn 8 triệu người dùng trên toàn thế giới [22]. Hệ thống của iNaturalist đã phát triển đến phiên bản 2.12 với 86.861 taxa, đạt độ chính xác trung bình 87.5% [23]. Tuy nhiên, một hạn chế quan trọng là độ chính xác cao chỉ đạt được với các loài phổ biến, trong khi các loài hiếm vẫn gặp khó khăn trong nhận diện [24].

Các nghiên cứu gần đây trong thị giác máy tính cho thực vật đã chứng minh những tiến bộ đáng kể. Nghiên cứu của August et al. (2020) sử dụng PlantNet trên dữ liệu Flickr cho thấy với điểm phân loại cao, độ chính xác nhận diện ở mức họ và ngành đều vượt 85%, trong khi nhận diện ở mức loài đạt khoảng 70% [24]. Hiệu suất được cải thiện đáng kể khi hình ảnh tập trung vào một loài thực vật thay vì là một phần của cảnh phức tạp. Wäldchen và Mäder (2018) trong tổng quan toàn diện về machine learning cho nhận diện loài dựa trên hình ảnh đã chỉ ra những thách thức cốt lõi trong lĩnh vực này, bao gồm sự biến đổi hình thái theo mùa, điều kiện ánh sáng và giai đoạn phát triển của thực vật [6]. Nghiên cứu này nhấn mạnh nhu cầu phát triển các mô hình chuyên biệt cho từng khu vực địa lý và nhóm loài cụ thể.

#### **1.2.4 Hệ thống hỏi đáp chuyên biệt cho thực vật**

Lĩnh vực hỏi đáp chuyên biệt cho thực vật đang trong giai đoạn phát triển với một số nghiên cứu đột phá. Zhou et al. (2021) đã phát triển phương pháp nhận diện bệnh cây trồng dựa trên multimodal deep learning, kết hợp thông tin hình ảnh và văn bản để đưa ra chẩn đoán và lời khuyên điều trị [13]. Nghiên cứu này mở ra hướng tiếp cận mới cho việc tích hợp kiến thức đa nguồn trong hệ thống hỏi đáp thực vật. Kolluri et al. (2024) đã đề xuất hệ thống nhận diện bệnh thực vật dựa trên multimodal learning, cho thấy rằng việc sử dụng thông tin đa nguồn (hình ảnh, văn bản mô tả, điều kiện môi trường) có thể cải thiện đáng kể độ chính xác so với các phương pháp chỉ dựa trên một phương thức duy nhất [14].

Yang et al. (2024) đã giới thiệu PLLaMa, một LLM mã nguồn mở đầu tiên được thiết kế chuyên biệt cho khoa học thực vật [15]. Mô hình này được đào tạo trên bộ dữ liệu khoa học thực vật lớn và cho thấy khả năng vượt trội trong việc trả lời các câu hỏi chuyên môn về thực vật so với các LLM tổng quát. Bên cạnh đó, các nghiên cứu về Retrieval-Augmented Generation (RAG) đã chứng minh tiềm năng lớn cho việc xây dựng hệ thống hỏi đáp chuyên biệt. Lewis et al. (2020) trong nghiên cứu tiên phong về RAG đã chỉ ra rằng việc kết hợp retrieval với generation có thể cải thiện đáng kể tính chính xác thực tế và khả năng cập nhật thông tin mà không cần đào tạo lại toàn bộ mô hình [25].

#### **1.2.5 Ứng dụng AI trong bảo tồn sinh học**

Việc ứng dụng AI trong bảo tồn sinh học đã chứng minh hiệu quả trong nhiều dự án quy mô lớn. Gillespie et al. (2024) đã phát triển mô hình Deepbiosphere sử dụng hơn 652.000 quan sát từ iNaturalist kết hợp với dữ liệu viễn thám để tạo ra bản đồ phân bố 2.221 loài thực vật tại California với độ phân giải chưa từng có [26]. Mô hình này đạt độ chính xác 89% trong nhận diện sự hiện diện của loài, vượt trội so với 27% của các phương pháp truyền thống. Nghiên cứu này đặc biệt có ý nghĩa khi mô hình có thể dự đoán chính xác 81.4% vị trí của cây redwood tại Redwood National Park và nắm bắt chính xác mức độ thiệt hại do cháy rừng Rim Fire năm 2013 với  $R^2=0.53$  [26]. Điều này chứng minh khả năng ứng dụng AI không chỉ trong nhận diện loài mà còn trong đánh giá tác động môi trường.

Loarie et al. (2017) trong nghiên cứu về thị giác máy tính của iNaturalist đã chỉ ra mô hình hợp tác giữa AI và cộng đồng có thể tạo ra vòng lặp cải thiện liên tục [27]. Mỗi quan sát và nhận diện được thêm vào hệ thống đều góp phần cải thiện mô hình, tạo ra một

hệ sinh thái học tập tự động. Điều này đặc biệt quan trọng đối với các khu vực có đa dạng sinh học cao như rừng Đà Nẵng, nơi sự tham gia của cộng đồng địa phương có thể cung cấp kiến thức bản địa quý giá.

Tuy nhiên, như Bartlett et al. (2023) đã lưu ý trong nghiên cứu về nấm Hebeloma, các hệ thống AI hiện tại vẫn gặp hạn chế với các loài hiếm hoặc có đặc điểm hình thái phức tạp [28]. Điều này nhấn mạnh tầm quan trọng của việc phát triển các hệ thống chuyên biệt cho từng khu vực sinh thái, kết hợp cả kiến thức khoa học hiện đại và truyền thống địa phương để đạt được hiệu quả tối ưu trong bảo tồn đa dạng sinh học.

#### **1.2.6 Mục tiêu nghiên cứu**

Nghiên cứu này hướng đến giải quyết các thách thức đã nêu thông qua việc phát triển hệ thống trả lời câu hỏi về thực vật thông minh được thiết kế riêng cho khu vực Đà Nẵng. Các mục tiêu chính tập trung vào việc tạo ra giải pháp thực tế, thu hẹp khoảng cách giữa công nghệ trí tuệ nhân tạo (AI) tiên tiến và nhu cầu thông tin về thực vật trong thế giới thực.

**Mục tiêu 1: Hệ thống phân loại thực vật hiệu chính xác:** Phát triển hệ thống phân loại thực vật đạt độ chính xác top 5 trên 85% đối với các loài thực vật rừng tại Đà Nẵng.

**Mục tiêu 2: Xử lý câu hỏi tiếng Việt:** Vận dụng các công nghệ xử lý ngôn ngữ tự nhiên cho tiếng Việt, xử lý hiệu quả sự phức tạp về mặt ngôn ngữ của tiếng Việt.

**Mục tiêu 3: Xây dựng cơ sở dữ liệu toàn diện cho khu vực:** Phát triển cơ sở kiến thức toàn diện bao gồm các loài thực vật ở Đà Nẵng, tích hợp danh pháp khoa học, mô tả hình thái, đặc điểm sinh thái và kiến thức sinh thái truyền thống. Cơ sở hạ tầng kiến thức này sẽ hỗ trợ cả nhiệm vụ cung cấp thông tin toàn diện cho các loài đã xác định.

**Kết quả và tác động dự kiến :** Việc thành công đạt được các mục tiêu kể trên sẽ xây dựng nền hệ thống hỏi đáp thông minh được thiết kế riêng cho hệ sinh thái rừng tại Đà Nẵng. Hệ thống này sẽ đóng vai trò là nền tảng hỗ trợ nghiên cứu đa dạng sinh học, giáo dục môi trường và cải thiện công tác bảo tồn sinh thái rừng Đà Nẵng. Ngoài các ứng dụng thực tế ngay lập tức, nghiên cứu này sẽ chứng minh các phương pháp tiếp cận một cách có hệ thống để phát triển các hệ thống thông tin thực vật phù hợp về mặt văn hóa và ngôn ngữ, có thể được điều chỉnh cho các khu vực khác đang phải đổi mới với những thách thức tương tự trong việc tiếp cận thông tin và lập kế hoạch bảo tồn.

## CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

### 2.1. Kiến trúc Deep Learning cho nhận diện thực vật

#### 2.1.1. ConvNeXt V2

ConvNeXt V2, được phát triển bởi Woo et al. (2023), đánh dấu một bước tiến đáng kể trong thế giới mạng nơ-ron tích chập (CNN) và là một phần quan trọng của cuộc cách mạng hiện đại hóa CNN [29]. Thay vì chỉ tập trung vào cải tiến kiến trúc đơn lẻ như các nghiên cứu trước đây, ConvNeXt V2 đưa ra một cách tiếp cận toàn diện hơn: đồng thời tối ưu hóa kiến trúc và tích hợp hiệu quả với phương pháp học tự giám sát, cụ thể là thông qua masked autoencoder (MAE). Mấu chốt nằm ở việc nó giải quyết được một thách thức lớn mà CNN đã đối mặt bấy lâu nay: làm thế nào để kết hợp tốt với các phương pháp tự học hiện đại.

Sự phát triển từ CNN truyền thống đến ConvNeXt V2 cho thấy một sự dịch chuyển đáng kể: từ việc thiết kế kiến trúc dựa trên trực giác sang một cách tiếp cận có nền tảng khoa học vững chắc hơn. Trong khi ResNet tập trung xử lý vấn đề vanishing gradient bằng residual connections [30], ConvNeXt V2 tiếp nối bằng cách mạnh dạn đưa các nguyên lý thiết kế từ Vision Transformers vào CNN, và thậm chí còn tiến xa hơn với việc giới thiệu lớp Global Response Normalization (GRN). GRN là một đột phá quan trọng, giúp tăng cường "sức cạnh tranh" giữa các kênh đặc trưng và ngăn chặn hiện tượng suy giảm đặc trưng (feature collapse), nơi thông tin bị mất đi do các kênh trở nên quá giống nhau.

##### 2.1.1.1. Cơ chế Global Response Normalization và ý nghĩa lý thuyết

Lớp GRN trong ConvNeXt V2 hoạt động dựa trên nguyên lý chuẩn hóa thích ứng. Về cơ bản, nó điều chỉnh tỷ lệ của mỗi kênh đặc trưng (channel) dựa trên mức độ quan trọng tương đối của kênh đó so với tổng thể các kênh trong cùng một feature map. Công thức toán học có thể biểu diễn như sau:

$$GRN(X) = \gamma * \left( \frac{X}{\|X\|_2 + \varepsilon} \right) + \beta$$

Ở đây,  $X$  là tensor đầu vào,  $\gamma$  và  $\beta$  là các tham số mà mô hình sẽ học được,  $\|X\|_2$  là chuẩn L2 của vector đặc trưng, còn  $\varepsilon$  là một hằng số nhỏ để tránh chia cho zero. Thiết kế

này mang ý nghĩa lý thuyết quan trọng: nó trang bị cho mô hình khả năng tự động điều chỉnh và làm nổi bật các kênh có mức độ kích hoạt cao, tạo ra một dạng cơ chế "chú ý" (attention) ngầm định.

Trong bối cảnh phân loại thực vật, GRN đóng vai trò đặc biệt quan trọng. Nó giúp mô hình nhận diện và tập trung vào những đặc trưng mang tính phân biệt cao nhất giữa các loài. Chẳng hạn, khi cần phân biệt giữa các loài thực vật có hình dạng lá tương tự nhưng khác nhau về kết cấu bề mặt, GRN có thể tăng cường tín hiệu từ các kênh đặc trưng liên quan đến vân lá (texture patterns), đồng thời giảm bớt sự ảnh hưởng của các kênh liên quan đến hình dạng tổng thể (shape).

---

### **Algorithm 1 Pseudocode of GRN in a PyTorch-like style.**

---

```
# gamma, beta: learnable affine transform parameters
# X: input of shape (N,H,W,C)

gx = torch.norm(X, p=2, dim=(1,2), keepdim=True)
nx = gx / (gx.mean(dim=-1, keepdim=True)+1e-6)
return gamma * (X * nx) + beta + X
```

---

Hình 2-1: Thuật toán Global Response Normalization

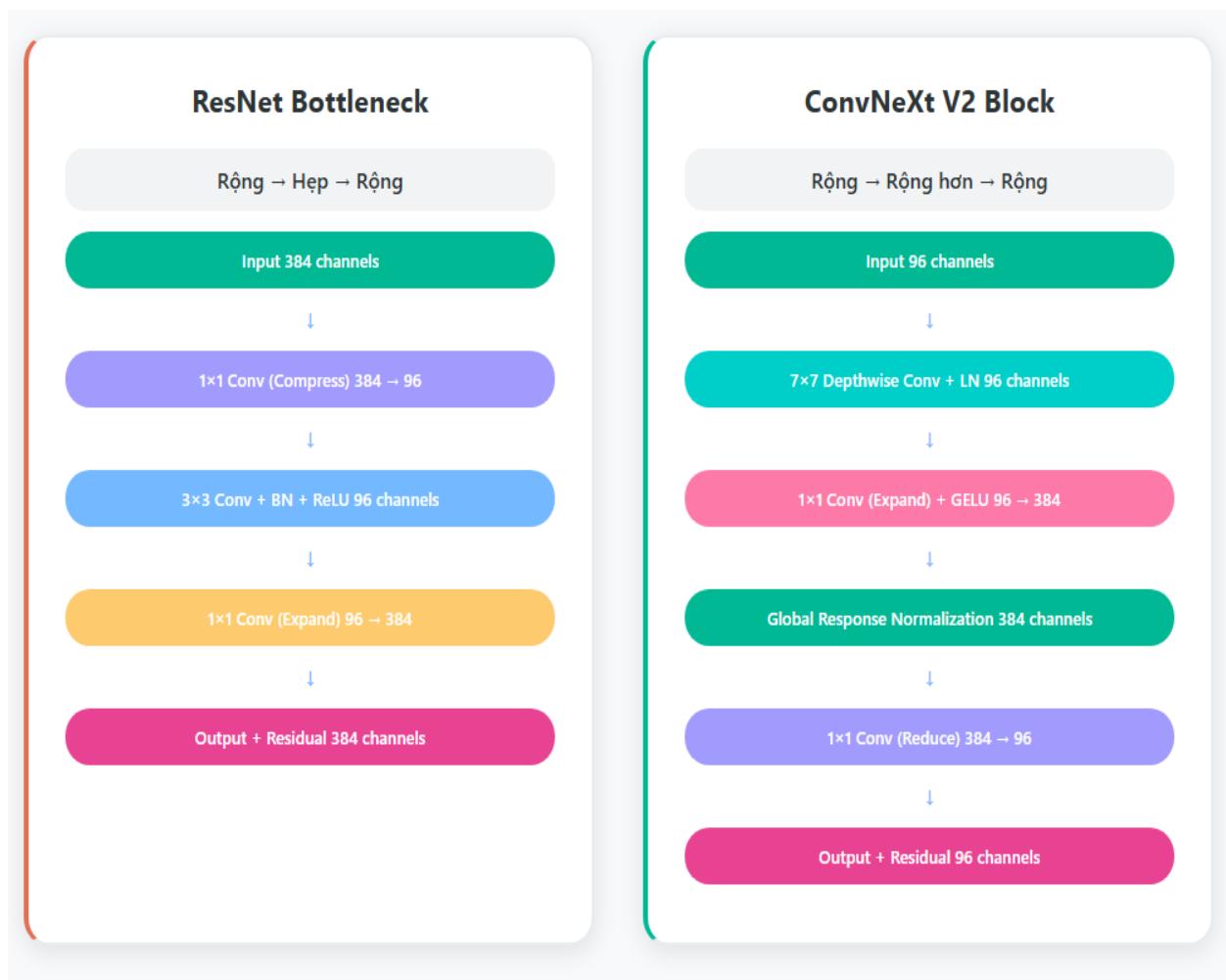
#### **2.1.1.2. Thiết kế bottleneck đảo ngược và ưu thế trong trích xuất đặc trưng**

Một trong những cải tiến then chốt khác của ConvNeXt V2 là việc áp dụng thiết kế inverted bottleneck (nghịch đảo bottleneck), một ý tưởng hoàn toàn khác biệt so với bottleneck truyền thống của ResNet. Trong khi ResNet thu hẹp chiều dữ liệu trước khi thực hiện phép tích chập 3x3 (từ "rộng" sang "hẹp" rồi lại "rộng"), ConvNeXt V2 lại bắt đầu bằng cách mở rộng chiều dữ liệu (từ "rộng" sang "rộng hơn" rồi lại "rộng"). Thiết kế này mang đến những lợi thế lý thuyết quan trọng:

- **Bảo toàn luồng thông tin:** Việc không thu hẹp chiều dữ liệu ngay từ đầu mỗi khối (block) giúp ConvNeXt V2 bảo toàn tối đa lượng thông tin được truyền từ các lớp trước. Điều này đặc biệt có ý nghĩa trong các tác vụ phân loại thực vật đòi hỏi độ tinh vi cao, nơi mà những chi tiết nhỏ như vân lá hay hình dạng răng cưa của lá cũng có thể là yếu tố then chốt để xác định loài.
- **Tăng cường tương tác đặc trưng:** Bằng cách mở rộng chiều dữ liệu, ConvNeXt V2 tạo điều kiện cho nhiều "đơn vị thần kinh" tương tác trong các lớp trung gian,

từ đó giúp mô hình học được những mẫu phức tạp hơn. Đối với thực vật học, điều này có nghĩa là mô hình có thể nắm bắt được các mối quan hệ phức tạp giữa các đặc trưng hình thái khác nhau của cây, chẳng hạn như mối liên hệ giữa hình dạng lá, cách sắp xếp hoa và cấu trúc thân.

- **Hiệu quả tính toán:** Dù thoát nhìn có vẻ việc mở rộng chiều dữ liệu sẽ làm tăng chi phí tính toán, thực tế thiết kế này lại tối ưu hơn đáng kể. Điều này có được là nhờ việc tận dụng các phép tích chập tách sâu (depthwise separable convolutions) và các hàm kích hoạt được tối ưu hóa, giúp giảm gánh nặng tính toán mà vẫn duy trì hiệu suất.



Hình 2-2: So sánh kiến trúc của ResNet và ConvNextV2

### **2.1.1.3 *Khả năng trích xuất đặc trưng cục bộ cho ứng dụng thực vật học***

ConvNeXt V2 thể hiện ưu thế vượt trội trong việc trích xuất các đặc trưng cục bộ, nhờ vào một số cơ chế thiết kế thông minh. Điểm hình là việc sử dụng kernel convolutions với kích thước lớn ( $7 \times 7$ ) trong lớp depthwise convolution. Điều này cho phép mô hình có một "vùng nhìn" (receptive field) rộng hơn, giúp nắm bắt các mẫu thực vật phức tạp một cách hiệu quả hơn, chẳng hạn như cách sắp xếp của lá kép hay cấu trúc chi tiết của cụm hoa.

- **Học đặc trưng phân cấp:** ConvNeXt V2 được thiết kế với bốn giai đoạn chính, mỗi giai đoạn xử lý thông tin ở một độ phân giải khác nhau. Ban đầu, các giai đoạn này tập trung nắm bắt những chi tiết tinh vi, có độ phân giải cao như kết cấu lá hay thông tin cạnh. Sau đó, các giai đoạn tiếp theo sẽ dần dần trừu tượng hóa thông tin, chuyển đổi chúng thành các khái niệm cấp cao hơn, ví dụ như cấu trúc tổng thể của cây và mối quan hệ không gian giữa các bộ phận. Cách tiếp cận này tạo ra một hệ thống đặc trưng phân cấp rất tự nhiên, hoàn toàn phù hợp với cách các nhà thực vật học tiến hành phân loại: từ các chi tiết vi mô đến hình thái vĩ mô.
- **Thiên hướng quy nạp không gian:** Không như Transformers thường phải tự học các mối quan hệ không gian từ đầu, CNN đã có sẵn "thiên hướng quy nạp không gian" (spatial inductive bias) thông qua các phép toán tích chập (convolution). ConvNeXt V2 khai thác hiệu quả đặc tính này, giúp mô hình nhanh chóng nắm bắt các mẫu không gian quan trọng trong hình ảnh thực vật, ví dụ như tính đối xứng, cấu trúc lặp lại và cách sắp xếp không gian của lá hoặc hoa.
- **Khả năng xử lý đa tỷ lệ:** Nhờ quy trình giảm mẫu (downsampling) tiến bộ và tổng hợp đặc trưng qua các giai đoạn, ConvNeXt V2 có thể xử lý hiệu quả hình ảnh thực vật ở nhiều tỷ lệ khác nhau cùng lúc. Đây là một lợi thế lớn, đặc biệt hữu ích khi xử lý các bức ảnh cây được chụp từ nhiều khoảng cách hoặc với các mức độ phóng to khác nhau – một thách thức thường gặp trong công việc thực địa của các nhà thực vật học.

### **2.1.1.4. *Transfer learning hiệu quả và khả năng thích ứng tốt***

Một trong những ưu điểm đáng chú ý nhất của ConvNeXt V2 khi ứng dụng trong thực vật học là khả năng transfer learning (học chuyển giao) hiệu quả từ các mô hình được tiền huấn luyện trên bộ dữ liệu ImageNet. Các nghiên cứu đã chỉ ra rằng ConvNeXt V2 đạt độ chính xác 88.9% trên ImageNet với chỉ 659 triệu tham số, thậm chí còn vượt

trội so với các phương pháp Transformer có cùng quy mô về hiệu quả tính toán (600.7 gigaflops so với 763.5 gigaflops của MViTV2) [29].

- **Khả năng chuyển đổi đặc trưng:** Các đặc trưng mà ConvNeXt V2 học được từ ImageNet - dù được huấn luyện trên các hình ảnh tự nhiên tổng quát - lại có khả năng chuyển đổi rất tốt sang miền thực vật học. Điều này là do các đặc trưng cấp thấp như cạnh, kết cấu và mẫu màu sắc có mối tương quan cao với các đặc trưng thực vật. Tương tự, các đặc trưng cấp cao hơn như hình dạng đối tượng và cách sắp xếp không gian cũng có liên quan trực tiếp đến hình thái thực vật, giúp mô hình dễ dàng thích nghi.
- **Giảm thiểu khoảng cách miền:** Khả năng học đặc trưng mạnh mẽ của ConvNeXt V2 đóng vai trò quan trọng trong việc thu hẹp "khoảng cách miền" (domain gap) giữa ImageNet và hình ảnh thực vật. Cụ thể, lớp GRN đặc biệt hiệu quả trong việc thích ứng các đặc trưng đã được tiền huấn luyện với miền dữ liệu mới, nhờ khả năng tự động điều chỉnh linh hoạt tầm quan trọng của các kênh đặc trưng dựa trên những mẫu hình thái đặc thù của thực vật.

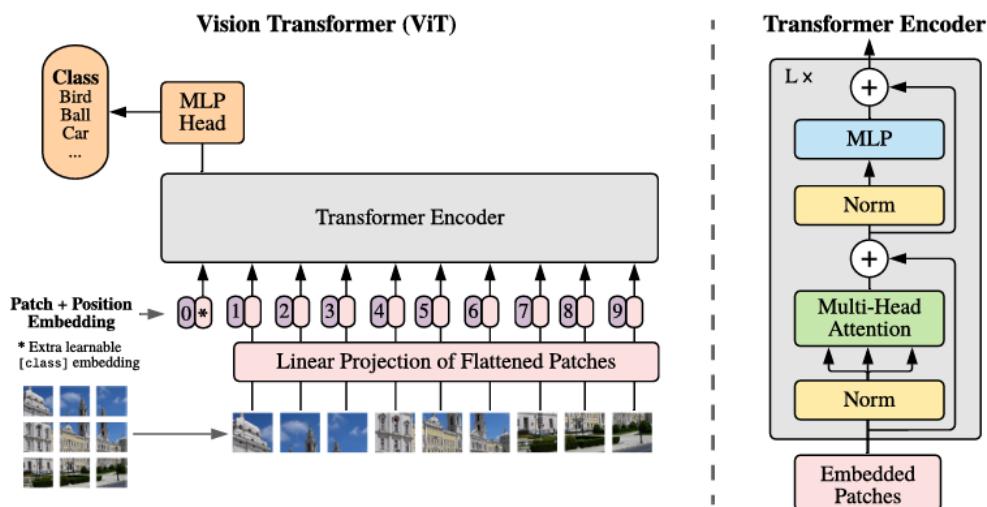
Tóm lại, ConvNeXt V2 là một bước tiến đột phá trong việc hiện đại hóa các kiến trúc CNN, đồng thời khẳng định tiềm năng ứng dụng mạnh mẽ của chúng trong lĩnh vực thực vật học. Với thiết kế kiến trúc tiên tiến - bao gồm lớp Global Response Normalization, inverted bottleneck và khả năng học đặc trưng phân cấp – ConvNeXt V2 có thể trích xuất và phân tích hiệu quả các đặc trưng hình thái tinh vi của thực vật ở nhiều mức độ khác nhau. Khả năng thích ứng linh hoạt khi chuyển đổi miền dữ liệu và hiệu quả tính toán vượt trội đã giúp ConvNeXt V2 trở thành một lựa chọn lý tưởng cho các ứng dụng thực tế, từ việc phân loại chính xác các loài đến hỗ trợ công tác bảo tồn và giám sát đa dạng sinh học.

### 2.1.2. Vision Transformers

Vision Transformers (ViT), được giới thiệu lần đầu bởi Dosovitskiy và cộng sự (2021), đại diện cho một bước tiến quan trọng trong lĩnh vực thị giác máy tính khi lần đầu tiên áp dụng thành công kiến trúc Transformer - vốn nổi bật trong xử lý ngôn ngữ tự nhiên - cho các bài toán xử lý ảnh [31]. Khác với các mạng nơ-ron tích chập (CNN) truyền thống vốn dựa trên các phép toán cục bộ, ViT có khả năng học các biểu diễn hình ảnh có ý nghĩa thông qua việc nắm bắt các mối quan hệ toàn cục trong ảnh. Điều này đặc

biệt phù hợp với thực vật học, nơi các đặc trưng phân loại quan trọng thường mang tính hình thái phức tạp và liên quan đến bối cảnh không gian tổng thể.

Cơ chế self-attention trong ViT cho phép mô hình tính toán trọng số tương quan giữa mọi cặp patch ảnh. Mỗi ảnh đầu vào được chia thành các patch cố định (ví dụ,  $16 \times 16$  pixel), sau đó các patch này được ánh xạ vào không gian vector và xử lý qua nhiều tầng Transformer. Quá trình này giúp ViT học được các biểu diễn giàu thông tin, không chỉ dựa trên đặc trưng cục bộ mà còn phụ thuộc vào ngữ cảnh toàn cục. Điều này có ý nghĩa lý thuyết sâu sắc trong thực vật học, khi các đặc trưng như sắp xếp lá, hình thái hoa, cấu trúc thân thường phụ thuộc mạnh vào quan hệ không gian giữa các bộ phận.



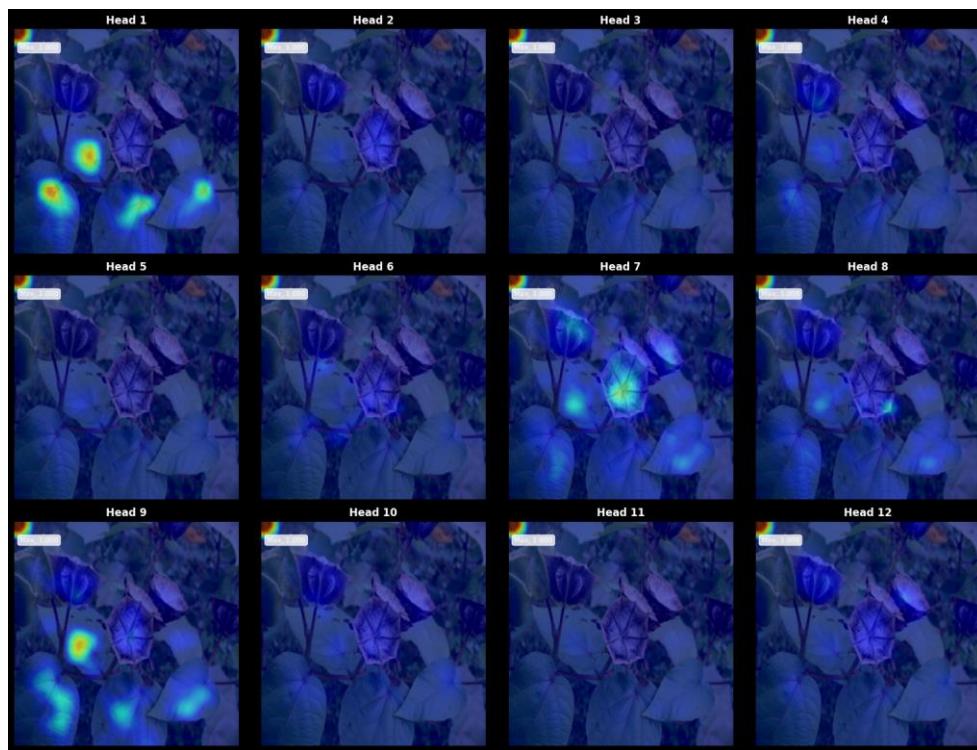
Hình 2-3: Kiến trúc Vision Transformer [31]

Qua các phiên bản cải tiến như DeiT (Data-efficient Image Transformer) [52] hoặc ViT-MAE [53], quy trình huấn luyện ViT đã được tối ưu hóa đáng kể, giúp tăng hiệu quả học biểu diễn từ dữ liệu không giám sát. Đặc biệt, việc áp dụng các kỹ thuật data augmentation phức tạp như color jittering, random crop, và geometric transformations giúp ViT trở nên bền vững hơn trước các biến thiên điều kiện chụp ảnh — vốn là thách thức phổ biến trong nhiếp ảnh thực địa cho ngành thực vật học.

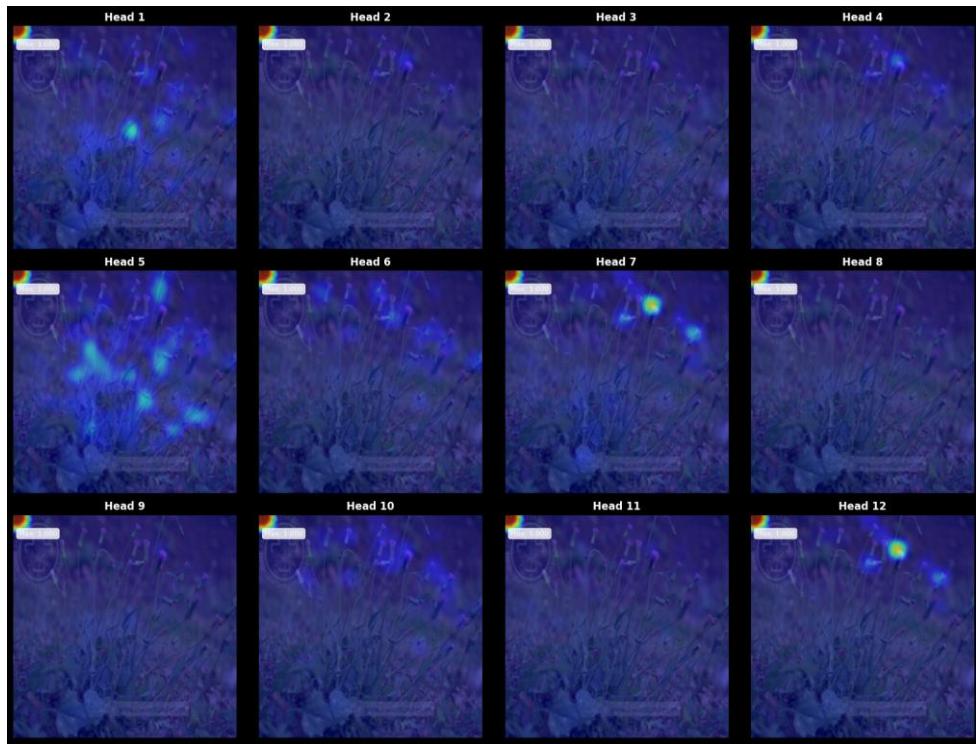
### 2.1.2.1. Cơ chế self-attention và hiểu ngữ cảnh toàn cục

Một trong những ưu điểm nổi bật của ViT so với CNN là khả năng học được các mối quan hệ không gian phức tạp giữa các bộ phận của cây. Self-attention trong ViT cho phép mô hình đồng thời học trọng số tương quan giữa mọi cặp vị trí trong ảnh, giúp nhận

diện các cấu trúc đối xứng, các mẫu sắp xếp phức tạp vốn khó khai thác bằng các phép toán cục bộ của CNN. Ví dụ, khi nhận diện một bông hoa đối xứng, ViT có thể đồng thời chú ý đến mối quan hệ giữa các cánh hoa, sự sắp xếp xoắn ốc của nhụy, và hình thái tổng thể. Đây là một ưu thế vượt trội so với CNN, vốn chỉ có khả năng học các đặc trưng cục bộ. Cơ chế multi-head attention cho phép ViT đồng thời học và nắm bắt nhiều loại đặc trưng khác nhau trong ảnh. Một attention head có thể tập trung vào màu sắc, một head khác vào kết cấu, và một head nữa vào hình học không gian. Sự phân tách học này rất có giá trị trong phân loại thực vật, khi các yếu tố như màu sắc hoa, kết cấu lá, mô hình sắp xếp của cành đều có thể là đặc trưng phân biệt quan trọng.



Hình 2-4: Minh họa attention của các head trong ViT với cây Tai mèo



Hình 2-5: Minh họa attention của các head trong ViT với cây Rau má tía

Từ minh họa trên có thể thấy, với mỗi lớp, sự kích hoạt cũng như hoạt động của các head là khác nhau. Trong khi với cây Tai mèo, head 1, 7 và 9 hoạt động đặc biệt mạnh mẽ, thì head 1, 5, 7, 12 lại chịu trách nhiệm chính cho cây Rau má tía. Sự phân công này chính là cốt lõi sức mạnh của cơ chế multi-head self-attention. Thay vì buộc một mô hình duy nhất phải học tất cả mọi thứ - từ kết cấu bề mặt, màu sắc, đến hình dạng tổng thể - ViT có thể cung cấp các "chuyên gia" (head) khác nhau để học các khía cạnh bổ sung cho nhau. Khả năng này đặc biệt quý giá trong phân loại thực vật, một lĩnh vực mà các đặc trưng phân biệt có thể rất đa dạng và tinh vi. Bằng cách tổng hợp thông tin từ nhiều head, ViT có thể xây dựng một biểu diễn tổng thể mạnh mẽ, tăng cường khả năng phân biệt các loài có hình thái phức tạp.

Ngoài ra, ViT tích hợp các positional encodings giúp mô hình hiểu được vị trí tương đối của các patch trong ảnh. Điều này giúp khắc phục hạn chế của các CNN truyền thống, vốn có tính bắt biến dịch chuyển nhưng lại không nắm bắt tốt các mối quan hệ không gian tương đối - một yếu tố quan trọng trong nhận diện cấu trúc cây. Trong thực vật học, thông tin này rất quan trọng để phân biệt các kiểu lá so le, lá đồi sinh, hoặc hoa

mọc chùm - những đặc trưng phân loại không thể nhận diện chính xác nếu bỏ qua thông tin không gian.

### **2.1.2.2. Điểm mạnh trong miền thực vật học và ứng dụng trong hệ thống**

Dựa trên nguyên lý hoạt động của cơ chế self-attention, Vision Transformers sở hữu các ưu điểm nổi trội cho bài toán về lĩnh vực thực vật học:

- **Khám phá mẫu hình thái không giám sát:** Các nghiên cứu đã chỉ ra rằng ViT có khả năng tự phát hiện các bộ phận cấu trúc của đối tượng [31]. Trong thực vật học, điều này cho phép mô hình học được các đặc trưng như mẫu vân lá, tính đối xứng của hoa, cấu trúc thân cây mà không cần chú thích chuyên gia.

- **Tính bát biến với điều kiện ảnh chụp:** ViT, khi được huấn luyện với augmentation thích hợp, phát triển khả năng bền vững với điều kiện chụp ảnh khác nhau - một yếu tố quan trọng khi làm việc với ảnh thực địa trong thực vật học, nơi các yếu tố như ánh sáng, góc chụp, mùa vụ thay đổi liên tục.

- **Khả năng khai quật hóa xuyên loài:** Nhờ khả năng học biểu diễn toàn cục của multi-head self-attention, ViT có thể khai quật hóa tốt qua các ranh giới loài, giúp phát hiện các đặc trưng tiến hóa hội tụ hoặc các mẫu hình thái chung. Điều này rất hữu ích trong các nghiên cứu về đa dạng sinh học và phát sinh loài.

Trong hệ thống được phát triển trong luận văn, ViT được fine-tune trên tập ảnh thực vật thu thập tại khu vực rừng Đà Nẵng. Quá trình fine-tune áp dụng các kỹ thuật augmentation đặc thù cho thực vật học, bao gồm crop tập trung vào lá hoặc cụm hoa, biến đổi màu sắc để tăng khả năng chịu thay đổi ánh sáng môi trường tự nhiên, và augmentation hình học để mô phỏng các góc chụp đa dạng trong thực địa.

Nhờ khả năng self-attention toàn cục và khả năng học biểu diễn phức tạp, ViT trong hệ thống cho thấy ưu thế trong việc nhận diện các đặc trưng hình thái tổng thể, hỗ trợ mạnh mẽ cho các trường hợp mà đặc trưng cục bộ chưa đủ phân biệt. Việc kết hợp ViT với ConvNeXt V2 trong kiến trúc Dual-Stream Fusion giúp tăng đáng kể độ chính xác phân loại cho các loài thực vật có hình thái biến thiên mạnh theo mùa hoặc giai đoạn phát triển.

Nhìn chung, Vision Transformers đại diện cho một nền tảng mạnh mẽ để học các biểu diễn hình ảnh phức tạp trong lĩnh vực thực vật học. Khả năng self-attention toàn cục,

tính bất biến với điều kiện chụp ảnh, và khả năng khám phá mẫu hình thái không giám sát khiến ViT trở thành một công cụ lý tưởng cho các ứng dụng phân loại học, bảo tồn, và nghiên cứu đa dạng sinh học. Với những đặc điểm trên, ViT không chỉ cung cấp một nền tảng mạnh mẽ cho các ứng dụng thực tế trong nhận diện thực vật, mà còn mở ra hướng nghiên cứu mới trong việc phân tích và khai thác các đặc trưng hình thái học phức tạp ở cấp độ hình ảnh.

### **2.1.3. Dual-Stream Fusion Architecture**

#### **2.1.3.1. Động cơ lý thuyết và giả thuyết bổ sung**

Việc phát triển kiến trúc Dual-Stream Fusion được thúc đẩy bởi quan sát rằng ConvNeXt V2 và Vision Transformers sở hữu những điểm mạnh khác biệt và có khả năng hỗ trợ cho nhau trong việc hiểu và phân tích hình ảnh thực vật.

Nghiên cứu của Zhang et al. (2023) đề xuất mô hình lai kết hợp CNN và Transformer để nhận diện cỏ dại. Mô hình đạt độ chính xác 96.08% trên bộ dữ liệu DeepWeeds, cao hơn đáng kể so với 89.43% của Vision Transformer thuần. [32].

Dựa trên nền tảng này, kiến trúc được đề xuất trong luận văn mở rộng ý tưởng kết hợp các mạng có triết lý thiết kế khác biệt, nhằm tận dụng tối đa tính hỗ trợ giữa các dòng đặc trưng.

Giả thuyết cốt lõi của kiến trúc cho rằng ConvNeXt V2, với các thiên hướng quy nạp như tính cục bộ, bất biến dịch chuyển, xử lý phân cấp, có khả năng học tốt các đặc trưng hình thái chi tiết. Trong khi đó, Vision Transformers, nhờ vào attention toàn cục, khả năng hiểu biết ngữ nghĩa và tính bất biến góc nhìn, có thể học được các đặc trưng ngữ cảnh giàu khái quát. Hai dòng đặc trưng này được cho là hỗ trợ lẫn nhau thay vì dư thừa.

Nhận định này không chỉ có cơ sở trong các nghiên cứu về mạng nơ-ron mà còn tương đồng với các chuyên gia thực vật học: họ thường kết hợp giữa phân tích hình thái chi tiết và nhận diện toàn cục khi xác định loài thực vật.

#### **2.1.3.2. Thiết kế kiến trúc của Fusion Block**

Một điểm nổi bật trong thiết kế là việc sử dụng ba nhánh song song để xử lý các đặc trưng đầu vào đã kết hợp từ ConvNeXt V2 và Vision Transformer, nhằm tối ưu hóa khả năng khai thác thông tin.

Khối Dual-Stream Fusion Bloc được thiết kế để kết hợp hai nguồn đặc trưng: một từ ConvNeXt (giàu thông tin cục bộ), một từ ViT (giàu thông tin toàn cục). Đầu tiên, hai đặc trưng được nối lại và đưa qua các lớp Conv 1x1 để làm đồng nhất chiều kênh và chuẩn hóa biểu diễn. Từ đây, ba nhánh xử lý song song được triển khai:

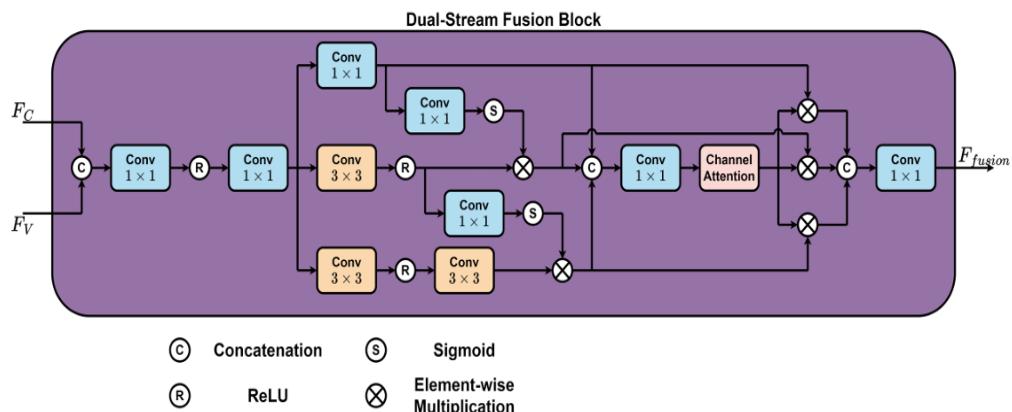
- **Nhánh 1 (refinement):** Sử dụng chuỗi các Conv 1x1 và sigmoid để sinh ra mặt nạ attention giúp điều chỉnh đặc trưng đầu vào.
- **Nhánh 2 (local enhancement):** Sử dụng các Conv 3x3 + ReLU để tăng cường tương tác cục bộ, sau đó nhân với mặt nạ từ nhánh 1.
- **Nhánh 3:** Tiếp tục tinh chỉnh đặc trưng bằng chuỗi Conv 3x3, sau đó được điều tiết bằng mặt nạ sigmoid từ nhánh 2.

Ba đặc trưng từ các nhánh này được nối lại và đưa qua attention theo kênh (ChannelAttention), sau đó nhân lại với từng nhánh tương ứng để làm nổi bật các vùng quan trọng. Cuối cùng, đầu ra hợp nhất được tạo ra bằng một lớp Conv 1x1 và đưa đến phân loại.

Quá trình kết hợp đặc trưng trong Fusion Block có thể được mô tả theo công thức:

$$F_{fusion} = \alpha_1 \cdot P_1(\text{Concat}(F_{conv}, F_{vit})) + \alpha_2 \cdot P_2(\text{Concat}(F_{conv}, F_{vit})) + \alpha_3 \cdot P_3(\text{Concat}(F_{conv}, F_{vit}))$$

trong đó  $\alpha_1, \alpha_2, \alpha_3$  là các trọng số có thể học được,  $P_1, P_2, P_3$  lần lượt đại diện cho ba nhánh xử lý, và  $F_{conv}, F_{vit}$  là các đặc trưng từ ConvNeXt V2 và Vision Transformer.



Hình 2-6: Kiến trúc khối Dual-Stream Fusion

Fusion Block còn tích hợp cơ chế attention trên các kênh (channel attention), đóng vai trò cân bằng động đóng góp từ các loại đặc trưng khác nhau. Channel attention học cách nhấn mạnh các kênh có khả năng phân biệt cao nhất cho các tác vụ phân loại thực vật cụ thể. Trong quá trình huấn luyện, cơ chế attention này sẽ khám phá ra các kết hợp nào của đặc trưng ConvNeXt V2 và Vision Transformer mang lại nhiều thông tin nhất cho việc phân biệt giữa các loài thực vật.

#### **2.1.3.3. Ưu thế lý thuyết cho phân loại thực vật tinh vi**

Kiến trúc Dual-Stream Fusion mang lại nhiều ưu thế lý thuyết quan trọng trong bài toán phân loại thực vật tinh vi. Trước hết, việc kết hợp các đặc trưng cục bộ phân cấp của ConvNeXt V2 với các đặc trưng ngữ cảnh toàn cục của Vision Transformer cho phép mô hình xử lý thông tin thực vật ở nhiều thang bậc, từ chi tiết vi mô của lá đến kiến trúc vĩ mô của cây.

Tiếp theo, mô hình có khả năng bền vững cao với các biến đổi hình thái trong loài - một đặc điểm rất thường gặp ở thực vật do ảnh hưởng của môi trường, giai đoạn phát triển hoặc đa dạng di truyền. Khi một luồng không nắm bắt được thông tin phân biệt do biến đổi hình thái bất thường, luồng còn lại có thể bù đắp.

Bên cạnh đó, sự kết hợp giữa các loại đặc trưng bổ sung giúp tăng cường sức mạnh phân biệt, đặc biệt hữu ích trong việc phân biệt các loài thực vật gần nhau về mặt hình thái. Các cơ chế attention còn cho phép mô hình triển khai động các loại "chuyên môn" khác nhau tùy thuộc vào đặc điểm của ảnh đầu vào, một cách tiếp cận gần gũi với phương pháp mà các chuyên gia thực vật học con người thường áp dụng trong thực tiễn.

Về hiệu suất, kiến trúc Dual-Stream Fusion được kỳ vọng sẽ cải thiện độ chính xác top-1 và top-k thông qua biểu diễn đặc trưng phong phú hơn, xử lý tốt hơn các trường hợp biên và ảnh thực vật thách thức, đồng thời tăng khả năng khái quát hóa cho các loài mới.

Cuối cùng, kiến trúc được thiết kế với tính mô-đun cao, giúp mở rộng dễ dàng khi áp dụng cho các tập dữ liệu thực vật quy mô lớn hơn hoặc khi chuyển sang các miền thực vật khác nhau (cây thân gỗ, cây thảo mộc, thực vật thủy sinh), thông qua tinh chỉnh đặc thù miền của các thành phần Fusion Block.

## 2.2. Metric learning và ProxyNCA loss

### 2.2.1 Lý thuyết phân loại dựa trên embedding

Phân loại dựa trên embedding đánh dấu một sự thay đổi cơ bản trong cách tiếp cận bài toán nhận diện. Thay vì học trực tiếp các siêu mặt phân tách (decision boundaries) trong không gian đặc trưng ban đầu, phương pháp này tập trung vào việc học một không gian biểu diễn mới. Trong không gian này, các mẫu cùng lớp được kéo lại gần nhau, trong khi các mẫu khác lớp lại được đẩy ra xa. Điểm khác biệt chính là thay vì tìm siêu mặt phân tách cố định như SVM hay hồi quy logistic, embedding-based classification học một hàm biến đổi. Hàm này ánh xạ dữ liệu đầu vào sang một không gian metric, nơi khoảng cách giữa các điểm trực tiếp thể hiện mức độ tương đồng ngữ nghĩa (semantic similarity).

Trong phân loại truyền thống, mục tiêu chính là tìm một siêu mặt phân tách tối ưu giữa các lớp. Cách tiếp cận này có thể hoạt động hiệu quả khi có đủ dữ liệu huấn luyện và các lớp được phân biệt rõ ràng. Tuy nhiên, trong bối cảnh phân loại thực vật - một dạng phân loại fine-grained - bài toán trở nên phức tạp do sự tương đồng cao giữa các lớp và sự biến đổi lớn trong nội bộ mỗi lớp. Các phương pháp embedding-based khắc phục vấn đề này bằng cách học một không gian biểu diễn trong đó khái niệm "tương đồng" được mã hóa một cách trực tiếp thông qua các metric khoảng cách. Ngoài ra, khả năng mở rộng linh hoạt là một ưu điểm nổi bật của embedding-based. Khi thêm các lớp mới, phương pháp truyền thống yêu cầu thay đổi kiến trúc và huấn luyện lại toàn bộ mô hình. Ngược lại, embedding-based chỉ cần tính toán embedding cho các mẫu mới và áp dụng bộ phân loại dựa trên khoảng cách, chẳng hạn k-NN. Đặc tính này đặc biệt phù hợp với bài toán phân loại thực vật, nơi việc khám phá loài mới hay phân loài là một quá trình liên tục.

Cốt lõi của phân loại dựa trên embedding là học một không gian metric, nơi khoảng cách giữa các điểm thể hiện mức độ liên hệ ngữ nghĩa. Không gian này thường phức tạp hơn không gian Euclidean truyền thống, giúp thể hiện tốt hơn các mối quan hệ phân cấp và ngữ nghĩa của thực vật. Quá trình học embedding là một bài toán tối ưu hóa, nhằm tìm một hàm ánh xạ  $f: X \rightarrow R^d$  sao cho  $d(f(x_i), f(x_j)) < d(f(x_i), f(x_k))$  với  $x_i, x_j$  cùng lớp và  $x_k$  khác lớp. Việc lựa chọn metric rất quan trọng; khoảng cách Euclidean đơn giản, trong khi độ tương đồng cosine phù hợp hơn cho embedding chiều cao, tập trung vào hướng vector và phản ánh tốt đặc trưng hình thái. Hơn nữa, embedding-based có khả năng đặc biệt trong việc nắm bắt các mối quan hệ phân cấp tự nhiên của thực vật. Trong một không gian embedding được học tốt, các loài cùng chi sẽ

tạo thành các cụm gần nhau, và các chi cùng họ sẽ có khoảng cách nhỏ hơn so với các họ khác, qua đó phản ánh cấu trúc phân loại tự nhiên.

### 2.2.2. ProxyNCA loss

ProxyNCA (Proxy Neighborhood Component Analysis), giới thiệu bởi Movshovitz-Attias et al. (2017) [33], là một cải tiến quan trọng trong học metric, giải quyết chi phí tính toán cao của các phương pháp dựa trên cặp (pairwise) hoặc bộ ba (triplet). Thay vì so sánh trực tiếp các cặp điểm dữ liệu, ProxyNCA sử dụng các vector đại diện có thể học được (proxy) cho mỗi lớp. Các proxy này được cập nhật thông qua gradient descent cùng với trọng số của mạng. Về bản chất, hàm mất mát ProxyNCA khuyến khích mỗi embedding của mẫu  $x_i$  được kéo lại gần proxy của lớp đúng  $p_{ci}$  và đẩy ra xa khỏi proxy của các lớp sai, được thể hiện qua công thức:

$$L_{\text{ProxyNCA}} = - \sum_i \log \left( \frac{\exp(-d(f(x_i), p_{ci}))}{\sum_j \exp(-d(f(x_i), p_j))} \right)$$

Trong đó  $f(x_i)$  là embedding của mẫu  $x_i$ ,  $p_{ci}$  là proxy của lớp đúng của  $x_i$ , và  $d(\cdot, \cdot)$  là hàm đo khoảng cách (thường là khoảng cách Euclidean bình phương).

So với các phương pháp pairwise/triplet, ProxyNCA có độ phức tạp  $O(n \times c)$  thay vì  $O(n^2)$  hay  $O(n^3)$ , trong đó  $n$  là kích thước batch và  $c$  là số lớp. Đây là một lợi thế quan trọng cho các ứng dụng thực vật với số lượng loài lớn. Ngoài ra, quá trình huấn luyện với ProxyNCA thường hội tụ nhanh hơn và ổn định hơn, nhờ các proxy cung cấp mục tiêu học cố định.

Một trong những yếu tố quan trọng ảnh hưởng đến hiệu quả huấn luyện là cách chuẩn hóa và điều chỉnh hệ số của embedding và proxy. Thông thường, cả embedding  $f(x_i)$  và proxy  $p_j$  đều được chuẩn hóa L2, sau đó nhân với các hệ số  $\alpha$  và  $\beta$ , theo công thức:

$$\hat{f}(x_i) = \alpha \cdot \frac{f(x_i)}{\|f(x_i)\|_2}, \quad \hat{p}_j = \beta \cdot \frac{p_j}{\|p_j\|_2}$$

Việc chuẩn hóa này giúp embedding và proxy tập trung vào quan hệ hướng (directional similarity), từ đó cải thiện khả năng phân biệt giữa các lớp, nhất là trong các bài toán phân loại tinh vi như phân loại loài thực vật. Nghiên cứu thực nghiệm chỉ ra rằng lựa chọn  $\alpha = 3$ ,  $\beta = 3$  thường phù hợp với các tập dữ liệu fine-grained, trong khi với các tập dữ liệu quy mô lớn hoặc mất cân bằng mạnh, giá trị  $\alpha = 1$ ,  $\beta = 8$  có thể mang lại kết quả tốt hơn [34].

Đặc điểm gradient của ProxyNCA Loss cũng đóng vai trò quan trọng trong việc giúp mô hình học tốt với dữ liệu mất cân bằng. Cụ thể, gradient của loss đối với embedding có dạng:

$$\frac{\partial L}{\partial f(x_i)} = 2[f(x_i) - p_{y_i}] - 2 \sum_{j=1}^C w_{ij}[f(x_i) - p_j]$$

Trong đó  $w_{ij}$  là trọng số softmax cho từng proxy. Điều này có nghĩa là embedding của mỗi mẫu vừa được "kéo" về phía proxy của lớp đúng, vừa được " đẩ'y" ra khỏi các proxy khác, tạo ra một hành vi học tương phản một cách tự nhiên mà không cần kỹ thuật negative sampling như trong các phương pháp triplet loss.

Việc khởi tạo các proxy cũng cần được chú ý. Một phương pháp đơn giản và hiệu quả là khởi tạo các proxy từ phân phối chuẩn  $N(0, \sigma^2)$ , giúp đảm bảo các proxy ban đầu phân bố đều trong không gian embedding. Trong các bài toán phân loại thực vật, có thể cân nhắc khởi tạo proxy sao cho phản ánh mối quan hệ phân loại học: các loài cùng chi hoặc cùng họ có proxy khởi tạo gần nhau, giúp mô hình dễ dàng học được cấu trúc phân cấp tự nhiên của sinh giới.

Một thách thức khi áp dụng ProxyNCA Loss cho các bài toán dữ liệu không cân bằng là thiết kế batch huấn luyện sao cho các lớp hiếm vẫn có đủ mẫu trong batch. Nếu chỉ lấy mẫu ngẫu nhiên, có thể xảy ra trường hợp các lớp hiếm không có hoặc rất ít mẫu trong batch, khiến quá trình cập nhật proxy cho các lớp này bị hạn chế. Do đó, cần thiết kế chiến lược sampling cân bằng, đảm bảo mỗi batch có sự hiện diện đủ mạnh của các lớp ít gặp.

Không gian embedding học được từ ProxyNCA Loss có khả năng phản ánh rõ ràng cấu trúc phân loại tự nhiên. Thông thường, các loài cùng chi sẽ tạo thành các cụm gần nhau trong không gian embedding trong khi các họ khác nhau có khoảng cách embedding

xa hơn. Đặc tính này rất hữu ích cho các ứng dụng như phân tích phát sinh loài (phylogenetic analysis), phát hiện lai tự nhiên, và hỗ trợ chiến lược bảo tồn đa dạng sinh học. Ngoài ra, embeddings phản ánh các đặc trưng cốt lõi của loài, trong khi vẫn cho phép linh hoạt với các biến dị tự nhiên (do di truyền, môi trường, mùa vụ). Đặc biệt, ProxyNCA Loss có khả năng mở rộng dễ dàng khi phát hiện loài mới: chỉ cần thêm một proxy mới cho loài đó mà không cần thay đổi kiến trúc mạng, đồng thời có thể tận dụng embedding đã học từ các loài gần gũi để hỗ trợ học nhanh hơn cho loài mới.



Hình 2-7: Mô tả quá trình cập nhật của Proxy NCA loss trong bài toán thực vật học

Trong hệ thống hỏi đáp thực vật mà luận văn đề xuất, embedding-based retrieval đóng vai trò quan trọng trong module tiền xử lý ảnh. Thông qua việc so sánh embedding của ảnh gửi đến với embedding của tập ảnh chuẩn, hệ thống có thể xác định liệu ảnh có thuộc miền (in-domain) hay không. Chỉ những ảnh in-domain mới được chuyển tiếp đến module phân loại chi tiết.

Điều này giúp giảm thiểu sai số phân loại do các ảnh ngoài miền, đồng thời tăng độ tin cậy và khả năng mở rộng của hệ thống khi triển khai trong môi trường thực tế.

### **2.3. Tìm kiếm tương đồng và phát hiện mẫu ngoài phân phối (OOD)**

#### **2.3.1. FAISS và bài toán tìm kiếm tương đồng trong không gian nhiều chiều**

##### **2.3.1.1. Động cơ và bối cảnh bài toán**

Việc truy hồi hiệu quả các mẫu thực vật tương đồng từ các cơ sở dữ liệu lớn là một yêu cầu cốt lõi trong các hệ thống phân loại thực vật dựa trên học sâu. Ứng dụng này đóng vai trò quan trọng không chỉ trong bài toán nhận diện loài mà còn trong các tác vụ tiếp theo như phát hiện mẫu ngoài phân phối (out-of-distribution, viết tắt là OOD).

Các mô hình học sâu hiện đại thường biểu diễn hình ảnh thực vật dưới dạng các vector nhúng (embedding vector) có số chiều cao, dao động từ 512 đến 2048 chiều. Điều này đặt ra những thách thức lớn cho bài toán tìm kiếm tương đồng, do "lời nguyền của không gian nhiều chiều" (curse of dimensionality): khi số chiều tăng cao, thể tích của không gian tăng theo hàm mũ, khiến các phương pháp tìm kiếm tuyến tính truyền thống hoặc các cấu trúc chỉ mục cổ điển trở nên kém hiệu quả.

Đặc thù của cơ sở dữ liệu thực vật là quy mô rất lớn, có thể chứa hàng chục ngàn loài với nhiều mẫu đại diện cho mỗi loài. Bài toán đặt ra là: làm sao để trong thời gian hợp lý, có thể tìm được các mẫu gần nhất trong không gian embedding phục vụ cho các tác vụ nhận diện và phân tích sau này.

##### **2.3.1.2. Nguyên lý của tìm kiếm tương đồng trong không gian embedding**

Bài toán tìm kiếm tương đồng (similarity search) có thể được mô tả như sau:

- Cho một vector truy vấn  $q \in \mathbb{R}^d$ , và một tập hợp  $N$  vector embedding  $X = \{x_1, x_2, \dots, x_N\}$  với mỗi  $x_i \in \mathbb{R}^d$ .
- Nhiệm vụ là tìm kiếm ra  $k$  vector trong tập  $X$  gần nhất với  $q$ , theo một hàm khoảng cách  $d(\cdot, \cdot)$  nhất định:

$$NN_k(q) = \arg \min_k d(q, x_i), \quad i \in \{1, \dots, N\}$$

Trong các hệ thống embedding cho thực vật, các vector embedding thường được chuẩn hóa theo chuẩn L2. Khi đó, ta có thể sử dụng khoảng cách Euclidean hoặc độ tương đồng cosine. Hai phép đo này trong trường hợp chuẩn hóa L2 sẽ cho kết quả sắp hạng

tương đương, trong khi khoảng cách Euclidean lại dễ tính toán hơn trên phần cứng hiện đại.

Khoảng cách L2 giữa hai vector đã chuẩn hoá có thể viết lại như sau:

$$d_{L2}(q, x_i) = \|q - x_i\|_2 = \sqrt{2 - 2q^T x_i}$$

Điều này cho thấy khoảng cách L2 và độ tương đồng cosine có mối liên hệ đơn điệu. Việc này cho phép ta tối ưu bài toán tìm kiếm trên phương diện tính toán trong khi vẫn giữ được ý nghĩa hình học của các vector embedding.

### 2.3.1.3. FAISS và các chiến lược triển khai cho ứng dụng thực vật

FAISS (Facebook AI Similarity Search) [58] cung cấp một tập hợp các thuật toán và cấu trúc dữ liệu tối ưu cho bài toán tìm kiếm tương đồng trong không gian nhiều chiều. Đây là một thư viện rất mạnh, được thiết kế để xử lý tập dữ liệu rất lớn với độ trễ thấp.

Trong bài toán nhận diện thực vật, độ chính xác của phép tìm kiếm là yếu tố quan trọng, bởi vì quyết định phân loại loài có thể ảnh hưởng trực tiếp đến các hành động bảo tồn hay nghiên cứu. Do đó, ta ưu tiên sử dụng các phương pháp tìm kiếm chính xác (exact search) thay vì các phương pháp xấp xỉ.

Trong FAISS, module IndexFlatL2 là một lựa chọn phù hợp khi cần đảm bảo độ chính xác tuyệt đối: nó thực hiện tìm kiếm toàn bộ (exhaustive search), tính toán trực tiếp khoảng cách từ vector truy vấn đến toàn bộ các vector trong cơ sở dữ liệu.

Độ phức tạp tính toán của phương pháp này là  $O(Nxd)$  cho mỗi truy vấn (với N là số lượng vector trong cơ sở dữ liệu, d là số chiều của vector) - điều này hoàn toàn khả thi với các cơ sở dữ liệu tầm trung (hàng chục ngàn mẫu), đặc biệt khi tận dụng các thư viện toán học tối ưu và lệnh SIMD mà FAISS đã hỗ trợ.

Với cơ sở dữ liệu lớn, vấn đề bộ nhớ luôn là một thách thức. FAISS hỗ trợ cơ chế ánh xạ tệp vào bộ nhớ (memory-mapped files), cho phép các chỉ mục lớn hơn dung lượng RAM vật lý vẫn có thể được truy cập hiệu quả thông qua cơ chế phân trang.

Điều này đặc biệt hữu ích cho các ứng dụng như cơ sở dữ liệu quốc gia về thực vật, nơi ta cần lưu trữ embedding cho toàn bộ các loài trong một khu vực sinh thái hoặc một quốc gia.

Để đảm bảo độ nhất quán của các phép đo tương đồng, cần chú ý đồng bộ hóa quy trình tiền xử lý và chuẩn hóa giữa quá trình huấn luyện embedding và quá trình xây dựng truy vấn. Cụ thể, tất cả các vector embedding cần được chuẩn hóa L2  $\hat{e} = \frac{e}{\|e\|_2}$

Hơn nữa, FAISS hỗ trợ khả năng cập nhật chỉ mục động: có thể thêm các vector embedding mới (khi có loài mới, mẫu mới, hoặc sửa đổi phân loại học) mà không cần xây dựng lại toàn bộ chỉ mục. Đây là yếu tố rất quan trọng trong bối cảnh dữ liệu thực vật luôn được cập nhật liên tục.

### **2.3.2. Phát hiện mẫu ngoài phân phối (OOD) bằng phương pháp trung bình khoảng cách theo lớp**

#### **2.3.2.1. Động cơ và thách thức của bài toán**

Phát hiện mẫu ngoài phân phối (out-of-distribution detection - OOD) là một bài toán đặc biệt quan trọng trong các hệ thống phân loại thực vật triển khai thực tế. Trong môi trường ứng dụng thật, người dùng có thể gửi các ảnh chứa các loài không có trong cơ sở dữ liệu, có chất lượng ảnh thấp hoặc thậm chí không phải ảnh thực vật

Nếu hệ thống không phát hiện chính xác các trường hợp này, hậu quả có thể rất nghiêm trọng: nhận diện sai một loài hiếm, hoặc cung cấp thông tin sai lệch có thể dẫn đến những hành động sai trong công tác bảo tồn hoặc nghiên cứu khoa học.

Không giống như nhiều bài toán thị giác máy tính khác, bài toán OOD trong nhận diện thực vật có nhiều đặc thù riêng:

- Độ đa dạng nội bộ (intra-class variation) cao: cùng một loài thực vật có thể có hình thái khác nhau tùy theo giai đoạn phát triển, mùa vụ, điều kiện môi trường, hoặc thậm chí các biến thể tự nhiên.

- Ranh giới giữa các loài (inter-class boundaries) có thể không rõ ràng, do hiện tượng tiến hoá gần, hoặc lai tạp (hybridization).

- Yêu cầu cao về độ chính xác trong các ứng dụng bảo tồn, nơi quyết định sai có thể dẫn đến những hậu quả tiêu cực.

Các phương pháp OOD dựa trên ngưỡng khoảng cách đơn giản (ví dụ: kiểm tra xem khoảng cách từ ảnh truy vấn đến mẫu gần nhất có vượt quá ngưỡng hay không) thường hoạt động không tốt trong bối cảnh này. Lý do là:

- Mẫu gần nhất có thể là ngoại lệ (outlier) trong lớp của nó → khoảng cách lớn hơn bình thường.

- Những loài có đa dạng hình thái cao sẽ tự nhiên có khoảng cách lớn hơn → dễ bị nhận nhầm là OOD.

### 2.3.2.2. Trung bình khoảng cách theo lớp

Phương pháp trung bình khoảng cách theo lớp (class-based distance averaging) khắc phục những hạn chế trên bằng cách khai thác một giả thuyết quan trọng:

Các mẫu của cùng một loài (class) thường tạo thành một cụm chặt chẽ trong không gian embedding. Việc tính khoảng cách trung bình từ mẫu truy vấn đến nhiều đại diện của một lớp sẽ cho kết quả ổn định và đáng tin cậy hơn so với chỉ sử dụng khoảng cách đến một mẫu đơn lẻ.

Cụ thể, với embedding truy vấn  $q$  và một lớp  $C$  có  $n_c$  mẫu  $\{c_1, c_2, \dots, c_{n_c}\}$ , ta tính khoảng cách trung bình như sau:

$$\bar{d}(q, C) = \frac{1}{n_c} \sum_{i=1}^{n_c} d(q, c_i)$$

Phương pháp này giúp:

- Giảm nhiễu do ảnh chất lượng thấp, góc chụp bất lợi, hoặc đặc điểm bất thường của một mẫu đơn lẻ.
- Ổn định hóa quyết định khi xử lý các loài có hình thái đa dạng.

Thuật toán phát hiện mẫu ngoài phân phối (OOD) được thiết kế nhằm tận dụng thông tin cấu trúc lớp trong không gian embedding của ảnh thực vật. Thay vì dựa trên khoảng cách đến mẫu gần nhất (nearest neighbor) vốn dễ bị nhiễu bởi các ngoại lệ, thuật toán này sử dụng trung bình khoảng cách theo lớp để đưa ra quyết định một cách ổn định hơn.

Quy trình thuật toán gồm các bước chính như sau. Đầu tiên, hệ thống sử dụng FAISS để truy hồi  $k$  mẫu gần nhất trong không gian embedding so với ảnh truy vấn  $q$ . Giá trị  $k$  được chọn đủ lớn (thông thường từ 100 đến 200) nhằm đảm bảo rằng tập mẫu truy

hồi có đủ đại diện từ nhiều lớp khác nhau, thay vì chỉ tập trung vào các lớp phổ biến hoặc gần nhất.

Sau khi truy hồi, hệ thống kiểm tra tính đa dạng của tập mẫu bằng cách đếm số lượng lớp khác nhau có trong tập k mẫu. Nếu số lượng lớp chưa đạt ngưỡng tối thiểu yêu cầu (ví dụ, ít hơn 10 lớp), thuật toán sẽ tự động tăng k và lặp lại quá trình truy hồi. Quá trình này tiếp diễn cho đến khi đạt được mức độ đa dạng lớp mong muốn hoặc khi k chạm đến giá trị tối đa được thiết lập.

Khi đã có đủ mẫu từ nhiều lớp, thuật toán tiến hành tính toán trung bình khoảng cách từ ảnh truy vấn q đến các mẫu của từng lớp. Cụ thể, với mỗi lớp  $C_j$  có ít nhất một mẫu trong tập truy hồi, ta tính giá trị trung bình khoảng cách:

$$\bar{d}_j = \frac{1}{n_j} \sum_{i \in C_j} d(q, x_i)$$

Trong đó  $n_j$  là số mẫu của lớp j trong tập truy hồi, và  $d(q, x_i)$  là khoảng cách giữa ảnh truy vấn và mẫu  $x_i$ .

Tiếp theo, thuật toán xác định khoảng cách trung bình nhỏ nhất trong số các lớp:

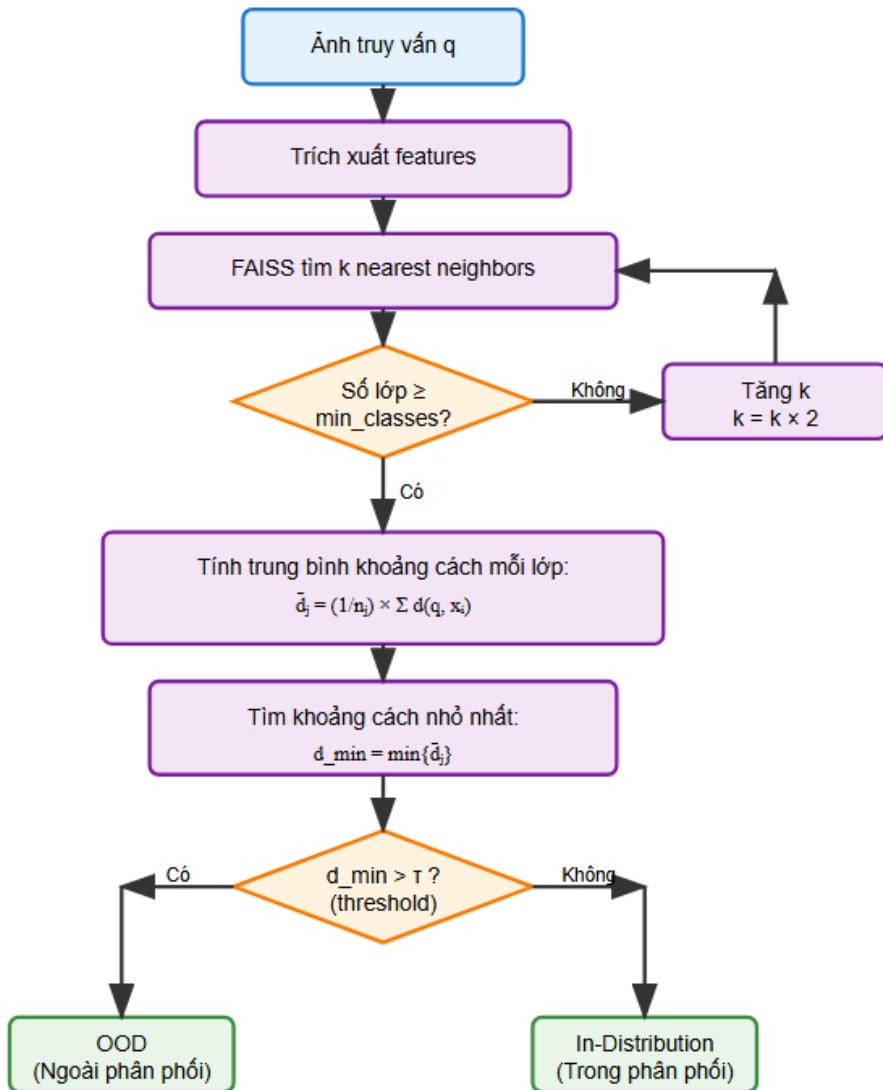
$$d_{min} = \min\{\bar{d}_1, \bar{d}_2, \dots, \bar{d}_m\}$$

Cuối cùng, giá trị  $d_{min}$  được so sánh với một ngưỡng  $\tau$ . Nếu  $d_{min} > \tau$ , ảnh truy vấn được phân loại là ngoài phân phôi (OOD); ngược lại, nếu  $d_{min} \leq \tau$ , ảnh được xem là nằm trong phân phôi (in-distribution).

Việc sử dụng trung bình khoảng cách theo lớp thay vì khoảng cách đến mẫu gần nhất giúp thuật toán giảm đáng kể phương sai trong quyết định. Theo định lý giới hạn trung tâm, trung bình của nhiều mẫu sẽ có phương sai giảm tỉ lệ nghịch với số lượng mẫu:

$$\text{Var}(\bar{d}) = \frac{\text{Var}(d)}{n}$$

### Thuật toán phát hiện OOD Class-based



Hình 2-8: Minh họa thuật toán phát hiện OOD

Nhờ vậy, thuật toán trở nên ít nhạy cảm hơn với các nhiễu ngẫu nhiên hoặc các mẫu ngoại lệ, vốn rất phổ biến trong dữ liệu thực vật do ảnh chụp không đồng nhất, dạng tự nhiên hoặc điều kiện môi trường khác nhau.

Bên cạnh đó, việc lấy trung bình trên nhiều mẫu giúp thuật toán chống chịu tốt hơn với các mẫu ngoại lệ nằm trong từng lớp. Khi một số mẫu trong lớp có thể là ngoại lệ (do lỗi chụp, sai nhãn, hoặc hình thái đặc biệt), trung bình hóa giúp làm giảm ảnh hưởng của các mẫu này đến quyết định cuối cùng.

Phương pháp cũng thích ứng tốt với độ đa dạng tự nhiên của các lớp. Trong thế giới thực vật, nhiều loài có hình thái rất đa dạng tùy theo mùa vụ, độ tuổi hay điều kiện sống. Việc dựa trên trung bình thay vì đơn lẻ một mẫu giúp thuật toán phản ánh đúng sự đa dạng hợp lệ của loài mà không dẫn đến nhận định sai là OOD.

Một ưu điểm quan trọng khác là hiệu quả tính toán. Các phương pháp phát hiện OOD dựa trên ước lượng mật độ trong không gian embedding thường yêu cầu mô hình hóa phân phối phức tạp và rất tốn tài nguyên, nhất là trong không gian embedding có số chiều cao. Ngược lại, phương pháp trung bình khoảng cách theo lớp có thể cài đặt đơn giản và chạy rất nhanh nhờ tận dụng FAISS - một thư viện tìm kiếm gần tối ưu trên không gian vector.

Trong bối cảnh ứng dụng nhận dạng thực vật và bảo tồn, phương pháp này còn mang lại nhiều lợi ích thiết thực. Đặc tính của cây cối là có sự biến thiên hình thái lớn theo mùa, tuổi và môi trường sống. Việc trung bình hóa theo lớp giúp hệ thống miễn nhiệm với các thay đổi hợp lệ này. Đồng thời, phương pháp còn giúp hạn chế nguy cơ nhầm lẫn giữa các loài hiếm và loài phổ biến - một vấn đề đặc biệt quan trọng trong bảo tồn. Với các loài hiếm, số mẫu trong tập huấn luyện thường rất ít, dễ bị lấn át bởi mẫu phổ biến. Việc ưu tiên quyết định bảo thủ (conservative), nghĩa là thiên về báo OOD khi không chắc chắn, giúp giảm nguy cơ nhận diện sai các loài hiếm - một nguyên tắc phù hợp với triết lý "nếu nghi ngờ thì từ chối nhận diện", rất được khuyến khích trong các ứng dụng bảo tồn.

Thuật toán còn có cơ chế chọn k thích ứng, tự động điều chỉnh giá trị k để đảm bảo đủ lớp đa dạng trong truy hồi. Bên cạnh đó, ngưỡng  $\tau$  cũng có thể được tinh chỉnh kỹ lưỡng trên các bộ dữ liệu đại diện để đạt được sự cân bằng mong muốn giữa độ nhạy (sensitivity) và độ đặc hiệu (specificity). Ngoài giá trị  $d_{min}$ , các chỉ số phụ như chênh lệch giữa  $d_{min}$  và lớp gần thứ hai, phuơng sai giữa các giá trị  $\bar{d}_j$ , hay số lượng lớp thực sự xuất hiện trong truy hồi cũng có thể được sử dụng để ước lượng định lượng độ tin cậy của từng quyết định.

Tuy vậy, phương pháp vẫn tồn tại một số hạn chế. Trong trường hợp các loài có hình thái gần nhau hoặc tiến hóa gần, embedding có thể chồng lấn, khiến việc phân biệt OOD trở nên khó khăn. Ngoài ra, nếu có sự dịch chuyển hệ thống (systematic shift) - chẳng hạn do thay đổi camera, điều kiện ánh sáng hoặc quy trình tiền xử lý - các khoảng cách

trong không gian embedding có thể bị sai lệch, ảnh hưởng đến hiệu quả của phương pháp. Cuối cùng, đối với các lớp có mẫu đại diện không đủ hoặc bị thiên lệch trong tập huấn luyện, giá trị trung bình khoảng cách có thể không phản ánh đúng cấu trúc lớp thực sự.

## **2.4. Xử lý ngôn ngữ Việt Nam cho hệ thống hỏi đáp về thực vật**

Phát triển hệ thống hỏi đáp về thực vật bằng tiếng Việt đòi hỏi sự hiểu biết sâu sắc về các nền tảng lý thuyết trong xử lý ngôn ngữ tự nhiên, đặc biệt là những thách thức đặc thù khi áp dụng vào lĩnh vực thực vật học. Phần này trình bày các cơ sở lý thuyết cần thiết để xây dựng hệ thống hỏi đáp hiệu quả cho ngôn ngữ Việt Nam trong chuyên ngành thực vật.

### **2.4.1. Lý thuyết ngữ nghĩa phân bố (Distributional Semantics)**

Lý thuyết ngữ nghĩa phân bố dựa trên nguyên lý cơ bản gọi là "distributional hypothesis", do John Rupert Firth nổi tiếng với câu nói: "You shall know a word by the company it keeps" [36]. Giả thuyết này cho rằng những từ có ý nghĩa tương tự sẽ xuất hiện trong những ngữ cảnh ngôn ngữ tương đồng.

Lý thuyết distributional semantics là một phương pháp định lượng nhằm hiểu ngữ nghĩa thông qua việc xây dựng không gian semantic, trong đó các từ được biểu diễn dưới dạng vector dựa trên các đặc tính phân bố trong ngôn ngữ [37]. Mục tiêu là đo được mức độ tương tự giữa các từ, cụm từ hoặc tài liệu thông qua khoảng cách trong không gian này.

Một bước phát triển quan trọng của lý thuyết này là ngữ nghĩa phân bố hợp thành, kết hợp giả thuyết phân bố với nguyên lý hợp thành, theo đó ý nghĩa của một cụm từ hay câu được xác định bởi ý nghĩa của các thành phần cấu tạo và cách chúng kết hợp cú pháp với nhau [38].

Trong hệ thống hỏi đáp về thực vật, việc hiểu ngữ nghĩa giúp xử lý các thuật ngữ phức hợp như "cây thuốc chữa ho" hay "lá xanh hình tim", nhờ đó hệ thống có thể hiểu đúng nghĩa tổng thể của câu hỏi dựa trên ngữ nghĩa các thành phần.

### **2.4.2. Không gian vector và tương tự ngữ nghĩa**

Vector embeddings là biểu diễn số học của các đơn vị ngôn ngữ nhằm nắm bắt đặc trưng ngữ nghĩa hoặc thuộc tính của chúng dưới dạng vector trong không gian liên tục, hỗ trợ việc so sánh và tìm kiếm hiệu quả [39].

Nguồn gốc của word embeddings dựa trên distributional semantics, trong đó các từ có ý nghĩa tương tự được biểu diễn bằng các vector gần nhau trong không gian đa chiều. Ví dụ, mô hình Word2Vec có thể mô phỏng các quan hệ ngữ nghĩa và cú pháp qua các phép toán vector như "Vua - Đàn ông + Phụ nữ = Nữ hoàng".

Để đo mức độ tương tự giữa các vector trong không gian semantic, cần sử dụng các hàm đo khoảng cách phù hợp. Cosine similarity được dùng phổ biến, nhưng có hạn chế khi độ lớn vector cũng mang thông tin ngữ nghĩa [40]. Euclidean distance đo khoảng cách thẳng giữa hai điểm, trong khi dot product kết hợp cả góc và độ dài vector.

Việc chọn hàm đo phù hợp rất quan trọng trong ứng dụng thực vật học, vì ảnh hưởng trực tiếp đến khả năng tìm kiếm các loài thực vật tương tự hoặc ứng dụng trong y học.

#### **2.4.3. Retrieval-Augmented Generation**

Retrieval-Augmented Generation (RAG), do Lewis và cộng sự phát triển năm 2020, là một phương pháp tinh chỉnh (fine-tuning) tổng quát giúp các mô hình ngôn ngữ lớn (LLM) kết nối với nguồn tri thức bên ngoài một cách linh hoạt [41]. RAG cho phép truy xuất thông tin thực tế từ kho dữ liệu bên ngoài, từ đó giúp mô hình ngôn ngữ lớn có cơ sở dữ liệu chính xác và cập nhật để tạo câu trả lời.

RAG kết hợp hai giai đoạn chính: truy xuất (retrieval) và sinh văn bản (generation). Tuy nhiên, hiệu quả của RAG phụ thuộc sâu vào chất lượng module truy xuất. Trong bối cảnh hỏi-đáp thực vật học – nơi độ chính xác và khả năng xử lý ngôn ngữ chuyên ngành là then chốt – chỉ dùng một phương pháp truy xuất đơn lẻ thường không đủ. Vì thế, hệ thống trong luận văn này áp dụng một cơ chế truy xuất kết hợp (hybrid retrieval), khai thác cả dense retrieval và sparse retrieval, rồi tổng hợp kết quả bằng Reciprocal Rank Fusion (RRF) [54].

Dựa trên giả thuyết phân phối (distributional hypothesis) đã đề cập tại Mục 2.4.1 rằng "You shall know a word by the company it keeps" [36], các biểu đạt có ý nghĩa gần nhau thường xuất hiện trong ngữ cảnh tương tự. Các kỹ thuật Dense Retrieval tận dụng nguyên lý này bằng cách sử dụng mô hình học sâu để ánh xạ truy vấn và tài liệu vào cùng một không gian vector liên tục (dense vector space). Trong không gian này, những câu hoặc cụm từ có nghĩa tương đồng sẽ được biểu diễn bởi các vector có khoảng cách nhỏ (hoặc độ tương đồng cosine cao).

Ưu điểm nổi bật của phương pháp này là khả năng nắm bắt ngữ nghĩa sâu sắc, xử lý tốt các hiện tượng ngôn ngữ như từ đồng nghĩa, hiện tượng paraphrasing, hay đa nghĩa ngữ cảnh. Điều này đặc biệt quan trọng trong các tình huống mà người dùng diễn đạt ý bằng những biểu thức khác với cách diễn đạt trong tài liệu gốc. Tuy nhiên, Dense Retrieval thường yêu cầu nhiều tài nguyên để huấn luyện và truy xuất, và đôi khi khó giải thích kết quả hơn so với các phương pháp dựa trên từ khóa.

Trái ngược với Dense Retrieval, Sparse Retrieval là phương pháp biểu diễn truy vấn và tài liệu bằng các vector thưa (sparse vectors), trong đó mỗi chiều tương ứng với một từ trong từ vựng và giá trị thể hiện mức độ quan trọng của từ đó trong văn bản. Các kỹ thuật điển hình như TF-IDF [55] và BM25 [56] thường được sử dụng để xây dựng các biểu diễn này.

Sparse Retrieval có ưu điểm là đơn giản, dễ triển khai và hiệu quả khi truy vấn chứa các thuật ngữ cụ thể, tên riêng, hay từ khóa hiếm – những trường hợp phổ biến trong các lĩnh vực như y học, thực vật học, hoặc pháp lý. Tuy nhiên, phương pháp này lại gặp khó khăn trong việc hiểu ngữ nghĩa sâu hoặc xử lý các truy vấn có hiện tượng đồng nghĩa, đa nghĩa, hay diễn đạt phức tạp, vì nó không xét đến ngữ cảnh của từ trong văn bản.

Nhằm kết hợp ưu điểm của cả hai phương pháp – khả năng hiểu ngữ nghĩa của Dense Retrieval và độ chính xác từ khóa của Sparse Retrieval – các phương pháp Hybrid Retrieval đã được phát triển và chứng minh hiệu quả vượt trội trong nhiều bài toán truy xuất văn bản [57].

Phương pháp này có thể được triển khai theo nhiều cách khác nhau, phổ biến nhất là late fusion (kết hợp kết quả từ hai hệ thống riêng biệt) hoặc joint indexing (xây dựng chỉ mục kết hợp cả hai dạng vector). Việc kết hợp này giúp cải thiện hiệu suất tổng thể: Dense component giúp hệ thống hiểu truy vấn ở cấp độ ngữ nghĩa, trong khi Sparse component đảm bảo không bỏ sót các từ khóa quan trọng. Trong đề tài này, thuật toán Reciprocal Rank Fusion được sử dụng để tổng hợp kết quả sau khi truy xuất. Đây là một thuật toán tính toán điểm quan trọng cho mỗi tài liệu truy xuất được, dựa trên thứ hạng của nó trong 2 danh sách kết quả (tài liệu nào được xếp hạng cao trong cả kết quả từ dense và sparse retrieval sẽ có điểm càng cao) với công thức như sau:

$$\text{RRF}(d) = \sum_{r \in R} \frac{1}{k + \text{rank}_r(d)}$$

Trong đó:

- d: là một tài liệu.
- R: tập hợp các danh sách xếp hạng tài liệu từ các phương pháp truy xuất khác nhau (dense, sparse)
- $\text{rank}_r(d)$ : thứ hạng của tài liệu d trong danh sách xếp hạng r.
- k: là một hằng số - thường được chọn là 60 – giúp điều hòa khoảng cách giữa điểm của các tài liệu xếp hạng gần nhau

Hybrid Retrieval đặc biệt hữu ích trong các hệ thống tìm kiếm tài liệu học thuật, y khoa hoặc đa ngôn ngữ, nơi vừa cần độ chính xác về mặt thuật ngữ, vừa cần khả năng bao phủ ngữ nghĩa linh hoạt. Nhờ tính thích nghi cao, nó có thể được hiệu chỉnh để ưu tiên một trong hai phương diện (ngữ nghĩa hoặc từ khóa) tùy vào mục tiêu ứng dụng cụ thể.

Bên cạnh Retrieval, một kỹ thuật quan trọng trong RAG là "prompt stuffing" - bổ sung thêm ngữ cảnh liên quan vào phần đầu của prompt nhằm hướng dẫn mô hình tập trung vào dữ liệu được cung cấp hơn là kiến thức đã được huấn luyện trước đó [42]. Trong hệ thống hỏi đáp thực vật, prompt stuffing cho phép thêm thông tin chi tiết về loài thực vật từ cơ sở tri thức, giúp câu trả lời về đặc điểm hình thái, phân bố hay ứng dụng y học trở nên chính xác hơn.

#### **2.4.4. Lý thuyết xử lý đa ngôn ngữ**

Xử lý ngôn ngữ tự nhiên cho tiếng Việt trong lĩnh vực thực vật học đặt ra nhiều thách thức đặc thù, đặc biệt khi hướng đến việc xây dựng biểu diễn ngữ nghĩa trong môi trường đa ngôn ngữ. Một trong những rào cản lý thuyết đáng chú ý là khái niệm “bất khả thông ước về mặt phân loại” (*taxonomic incommensurability*), vốn đề cập đến khó khăn trong việc so sánh và chuyển đổi giữa các hệ thống khái niệm khoa học khác nhau do sự khác biệt về cách sử dụng thuật ngữ, tiêu chí phân loại và phương pháp đo lường [43].

Trong thực tế, thách thức này thể hiện rõ qua sự chênh lệch giữa hệ thống phân loại khoa học quốc tế và cách gọi tên thực vật truyền thống trong văn hóa Việt Nam. Ví dụ, một loài cây có thể được biết đến rộng rãi dưới tên khoa học bằng tiếng Latin, trong khi

tại Việt Nam, nó lại được gọi bằng một loạt tên dân gian mang tính mô tả hoặc gắn với công dụng dân dã. Khoảng cách giữa hai hệ thống này gây khó khăn trong việc xây dựng các mô hình xử lý ngôn ngữ có khả năng hiểu và liên kết các biểu đạt khái niệm tương ứng.

Để giải quyết vấn đề đó, các mô hình embedding đa ngôn ngữ được phát triển dựa trên giả thuyết rằng các ngôn ngữ khác nhau có thể chia sẻ một không gian ngôn ngữ nghĩa chung, trong đó các khái niệm tương đồng sẽ được ánh xạ gần nhau bất kể ngôn ngữ gốc.

#### **2.4.5. Những thách thức lý thuyết đặc thù với tiếng Việt**

Tiếng Việt, với đặc điểm là một ngôn ngữ đơn lập và đơn âm tiết, đặt ra nhiều thách thức lý thuyết trong xử lý ngôn ngữ tự nhiên. Khác với các ngôn ngữ tổng hợp như tiếng Anh, nơi quan hệ ngữ pháp được biểu hiện rõ qua hình thái từ, tiếng Việt chủ yếu dựa vào trật tự từ và ngữ cảnh để truyền đạt ý nghĩa. Điều này khiến việc tách từ và biểu diễn ngữ nghĩa trở nên phức tạp hơn, đồng thời đòi hỏi các mô hình nhúng ngôn ngữ (embeddings) phải có khả năng nắm bắt thông tin vị trí và nhạy cảm với ngữ cảnh ở mức cao hơn.

Bên cạnh đặc điểm hình thái và cú pháp, một vấn đề mang tính lý thuyết khác là việc thích ứng các mô hình ngôn ngữ cho tiếng Việt - vốn được xếp vào nhóm ngôn ngữ ít tài nguyên. Lý thuyết về thích ứng mô hình ngôn ngữ đóng vai trò quan trọng trong việc phát triển các hệ thống NLP hiệu quả cho tiếng Việt. Theo đó, hai phương pháp chính thường được áp dụng là tiền huấn luyện liên tục và học chuyển giao liên ngôn ngữ, dựa trên giả thuyết rằng các cấu trúc ngôn ngữ cơ bản có thể được chia sẻ và chuyển giao giữa các ngôn ngữ khác nhau.

Hơn nữa, một thách thức lý thuyết mang tính đặc thù đối với tiếng Việt còn nằm ở tính hợp thành trong thuật ngữ thực vật. Nhiều tên gọi trong lĩnh vực này được cấu tạo từ sự kết hợp của các yếu tố như đặc điểm hình thái, màu sắc hoặc công dụng, ví dụ như "rau ngót" (rau + ngót) hay "cây thuốc Lào" (cây + thuốc + Lào). Do đó, để hiểu chính xác ý nghĩa của các thuật ngữ phức hợp này, mô hình cần học được cách kết hợp các thành phần ngữ nghĩa và suy diễn chúng trong từng ngữ cảnh cụ thể.

#### **2.4.6. Mô hình ngôn ngữ lớn (Large Language Models)**

Mô hình ngôn ngữ lớn (LLM) đại diện cho một bước đột phá trong xử lý ngôn ngữ tự nhiên, được xây dựng trên nền tảng kiến trúc Transformer được giới thiệu trong nghiên

cứu đột phá "Attention Is All You Need" năm 2017 [45]. Khác với các mạng neural tái phát (RNN) truyền thống xử lý từ theo tuần tự [46], Transformer có thể xem xét toàn bộ câu cùng lúc thông qua cơ chế self-attention, làm cho chúng hiệu quả hơn đáng kể trong việc nắm bắt các sắc thái của ngôn ngữ.

Lý thuyết cốt lõi của LLM dựa trên nguyên tắc "dự đoán từ tiếp theo": cho một lời nhắc văn bản từ người dùng, từ có khả năng xuất hiện cao nhất tiếp theo là gì. Kiến trúc Transformer bao gồm hai thành phần chính: bộ mã hóa (encoder) đọc và xử lý văn bản đầu vào, và bộ giải mã (decoder) sử dụng thông tin đã xử lý để tạo ra đầu ra. Cơ chế self-attention cho phép mô hình cân nhắc tầm quan trọng của mỗi từ trong câu tương đối với tất cả các từ khác, giúp duy trì hiểu biết toàn diện về toàn bộ ngữ cảnh.

Từ góc độ triển khai và sử dụng, các mô hình ngôn ngữ lớn có thể được phân thành hai loại chính: API-hosted và self-hosted, mỗi loại có những đặc điểm và ứng dụng riêng biệt phù hợp với các nhu cầu khác nhau.

Các mô hình API-hosted như Claude của Anthropic, GPT của OpenAI, và Gemini của Google đại diện cho phương pháp tiếp cận "mô hình như dịch vụ". Những mô hình này được lưu trữ và vận hành trên cơ sở hạ tầng đám mây của các công ty công nghệ lớn, người dùng truy cập thông qua các API với việc tính phí dựa trên lượng sử dụng (thường tính theo token). Ưu điểm chính của phương pháp này bao gồm khả năng tiếp cận ngay lập tức đến các mô hình tiên tiến nhất mà không cần đầu tư vào phần cứng, được cập nhật và bảo trì liên tục bởi các chuyên gia, và khả năng mở rộng linh hoạt theo nhu cầu sử dụng. Tuy nhiên, hạn chế bao gồm chi phí vận hành liên tục, sự phụ thuộc vào kết nối internet, và hạn chế trong việc kiểm soát dữ liệu cũng như tùy chỉnh mô hình.

Ngược lại, các mô hình self-hosted như Qwen của Alibaba, Gemma của Google (phiên bản mã nguồn mở), và Llama của Meta cho phép tổ chức tự triển khai và vận hành mô hình trên cơ sở hạ tầng riêng. Phương pháp này cung cấp quyền kiểm soát hoàn toàn đối với dữ liệu và mô hình, khả năng tùy chỉnh sâu theo nhu cầu cụ thể, và tính độc lập không phụ thuộc vào dịch vụ bên ngoài. Tuy nhiên, self-hosted LLMs đòi hỏi đầu tư đáng kể vào phần cứng (đặc biệt là GPU chuyên dụng), chuyên môn kỹ thuật để triển khai và bảo trì, cũng như trách nhiệm tự cập nhật và tối ưu hóa mô hình. Đối với các ứng dụng đòi hỏi bảo mật cao hoặc xử lý dữ liệu nhạy cảm như thông tin cá nhân, tài chính, self-hosted models có thể là lựa chọn phù hợp hơn.

## CHƯƠNG 3: GIẢI PHÁP ĐỀ XUẤT

### 3.1. Tổng quan kiến trúc của hệ thống

Như đã đề cập trong Chương 1, việc xây dựng một hệ thống hỏi đáp về thực vật yêu cầu khả năng xử lý đa thể thức, bao gồm khả năng xử lý ảnh để nhận diện chính xác loài thực vật đang được đề cập và khả năng xử lý ngôn ngữ tự nhiên để nắm bắt được ý nghĩa của câu hỏi từ người dùng. Hơn nữa, như trong Chương 2 đã viết, các cách tiếp cận bằng học sâu – deep learning đã cho thấy những kết quả vượt trội trong cả tác vụ phân loại lẫn xử lý ngôn ngữ tự nhiên. Vì vậy, luận án này sẽ tận dụng và cải tiến các giải pháp học sâu để giải quyết các bài toán về nhận diện thực vật và truy xuất thông tin.

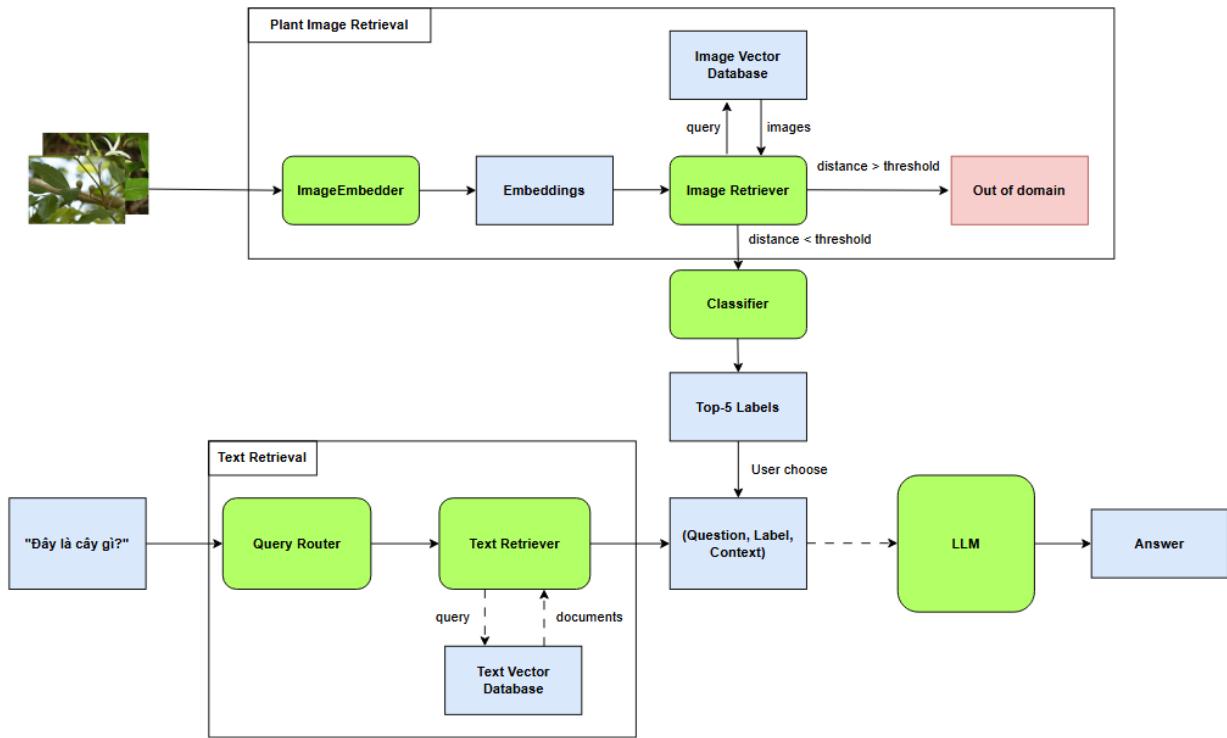
Để giải quyết các vấn đề đã nêu cũng như đạt được mục tiêu của đồ án, tôi đã phát triển FloraQA – một hệ thống hỏi đáp thông minh về thực vật rừng cho khu vực Đà Nẵng, với toàn bộ kiến trúc được minh họa trong Hình 3.1. Hệ thống bao gồm 4 thành phần chính, mỗi thành phần chịu trách nhiệm cho một nhiệm vụ chuyên biệt:

**Truy vấn ảnh thực vật:** Mô đun này quyết định 1 ảnh đầu vào của người dùng có chứa loài thực vật tồn tại trong danh sách các loài mà hệ thống hỗ trợ (in-domain) hay không. Để làm được điều này, trước tiên nó mã hóa ảnh đầu vào, sau đó thực hiện tìm kiếm tương đồng (Similarity search) trên 1 cơ sở dữ liệu vector của các ảnh đã được chuẩn bị sẵn.

**Phân loại thực vật:** Mô đun này phân loại các ảnh được đánh giá là in-domain vào các nhãn phù hợp.

**Truy xuất văn bản:** Mô đun này chịu trách nhiệm cho việc chuẩn bị các thông tin cần thiết để trả lời câu hỏi từ người dùng. Để làm được điều này, trước tiên nó mã hóa câu hỏi, sau đó thực hiện truy xuất các thông tin liên quan nhất từ cơ sở dữ liệu vector về văn bản đã được chuẩn bị sẵn bằng phép tìm kiếm tương đồng.

**Sinh văn bản:** Mô đun này tận dụng output từ mô đun truy xuất văn bản và phân loại thực vật để sinh ra câu trả lời phù hợp nhất đối với câu hỏi từ người dùng.



Hình 3-1: Sơ đồ tổng quan luồng hoạt động của hệ thống

Như được minh họa trong hình trên, hệ thống được thiết kế với ba luồng hoạt động chính, tự động kích hoạt dựa trên loại dữ liệu đầu vào từ người dùng:

- Ảnh:** Khi người dùng chỉ gửi hình ảnh, các ảnh này sẽ được xử lý tuân tự thông qua hai mô đun chính: mô đun truy xuất ảnh và mô đun phân loại ảnh. Hệ thống sẽ đưa ra danh sách *top-k* loài thực vật phù hợp nhất với nội dung hình ảnh. Người dùng sau đó sẽ lựa chọn một loài trong danh sách này để tiếp tục các thao tác kế tiếp. Cách tiếp cận phân loại theo *top-k* được đánh giá là phù hợp với các bài toán có số lượng lớp lớn, và cũng đang được áp dụng rộng rãi trên các nền tảng uy tín như PlantNet hay iNaturalist. Việc chỉ dựa vào phân loại *top-1* trong những trường hợp này thường không đảm bảo độ chính xác cần thiết. Từ góc độ trải nghiệm người dùng, *top-k* cũng mang lại tính linh hoạt cao hơn, cho phép người dùng tự chọn lựa loài mà họ cho là chính xác nhất dựa trên quan sát cá nhân.
- Văn bản:** Trong trường hợp người dùng chỉ nhập văn bản (câu hỏi) mà không kèm hình ảnh, hệ thống sẽ kích hoạt mô đun truy xuất văn bản và mô đun sinh văn bản. Trước tiên, câu hỏi sẽ được phân loại dựa trên các quy tắc đã định sẵn (ví dụ: câu

hỏi về một loài cụ thể, câu hỏi liên quan đến được tính, hay câu hỏi tổng quát). Sau đó, các ngữ cảnh liên quan sẽ được truy xuất và cung cấp cho mô đun sinh văn bản nhằm tạo ra câu trả lời chính xác, phù hợp với yêu cầu của người dùng.

- **Kết hợp:** Khi người dùng đồng thời gửi cả ảnh và văn bản, hệ thống ưu tiên xử lý phần hình ảnh trước để xác định tên loài thực vật. Thông tin tên loài sau đó sẽ được sử dụng làm ngữ cảnh cho mô đun truy xuất và sinh văn bản, từ đó tạo ra câu trả lời cuối cùng phù hợp với nội dung văn bản được nhập vào.

Trong các phần tiếp theo, báo cáo này sẽ cung cấp trình bày và giải thích chi tiết về quá trình xử lý của từng mô đun trong hệ thống

### 3.2. Truy xuất ảnh thực vật

Mô đun truy xuất hình ảnh thực vật là bước quan trọng đầu tiên trong FloraQA, có nhiệm vụ xác định xem ảnh đầu vào có chứa loài thực vật in-domain hay không. Nói cách khác, mô đun này hoạt động như một bộ lọc ảnh in-domain. Các chức năng chính của mô đun này bao gồm:

- **Mã hóa hình ảnh (Image Embedding):** tận dụng kiến trúc Dual-Stream Fusion (trình bày ở mục 2.1.3) để chuyển các ảnh thành các vector đặc trưng. Việc kết hợp khả năng trích xuất đặc trưng cục bộ của ConvNeXt V2 và hiểu ngữ nghĩa toàn cục của Vision Transformers giúp mô đun phù hợp cho nhiệm vụ này.

- **Tìm kiếm tương đồng (Similarity Search):** sau khi chuyển ảnh thành vector, mô đun tiến hành tìm kiếm độ tương đồng với các vector trong cơ sở dữ liệu vector đã xây dựng, áp dụng phương pháp trung bình khoảng cách theo lớp (class-based distance averagin) đã trình bày ở mục 2.3.2 nhằm phát hiện ảnh chứa các nội dung không được hỗ trợ (out-of-domain).

Đầu vào và đầu ra của mô đun này có thể được mô tả như sau:

- **Input:** ảnh (nên có độ phân giải cao, chứa 1 vật thể).

- **Output:** các ảnh được đánh giá là in-domain.

Trong bài toán truy xuất thông tin, chất lượng của bộ trích xuất đặc trưng thường đóng vai trò then chốt. Vì vậy, đồ án này tận dụng khói Dual-Stream Fusion đã được huấn luyện với hàm mất mát ProxyNCA để đảm nhiệm vai trò này.

Quá trình như sau:

- **Huấn luyện Dual-Stream Fusion** trên tập dữ liệu thực vật đã thu thập, sử dụng ProxyNCA loss giúp mô hình học được embedding có ý nghĩa, thể hiện tốt các đặc tính về sự biến thiên trong cùng một loài và sự tương đồng đa loài.

- **Xây dựng cơ sở dữ liệu vector:** Sau khi huấn luyện xong, tôi tận dụng mô hình này để tạo vector cho toàn bộ ảnh trong tập dữ liệu. Từ đó xây dựng cơ sở dữ liệu vector bằng FAISS để phục vụ cho việc so khớp ảnh.

- **So khớp ảnh đầu vào:** ảnh từ người dùng sẽ được ánh xạ thành vector cũng bởi mô hình trên, sau đó so khớp với cơ sở dữ liệu vector bằng phương pháp trung bình khoảng cách theo lớp, giúp xác định xem ảnh đó có thuộc một loài đã biết không.

### 3.3. Phân loại thực vật

Mô đun phân loại thực vật là thành phần tối quan trọng trong toàn bộ hệ thống FloraQA, đảm nhiệm việc nhận diện chính xác các loài thực vật từ ảnh. Mô đun này dự đoán tên khoa học của các ảnh chụp lá, hoa hoặc thân.

Để đạt hiệu quả cao, tôi sử dụng chiến lược phân loại tổng hợp (ensemble), kết hợp kết quả từ nhiều ảnh khác nhau của cùng một loài thực vật. Các chức năng chính của mô đun bao gồm:

- **Phân loại (Classification):** Sử dụng Dual-Stream Fusion kết hợp thêm 1 mạng neuron hai lớp để phục vụ cho việc phân loại.

- **Tổng hợp kết quả (Ensemble):** Nếu có nhiều ảnh, mô đun sẽ bắt đầu khâu hậu xử lý kết quả để tổng hợp kết quả phân loại của từng ảnh đơn lẻ dựa trên độ tin cậy, sau đó chọn ra top-k lớp cao nhất.

Đầu vào và đầu ra của mô đun này có thể được mô tả như sau:

- **Input:** ảnh thực vật in-domain, mỗi ảnh chỉ nên chứa một bộ phận của một loài thực vật.

- **Output:** vector chứa độ tin cậy của top-k lớp cao nhất.

Mô đun phân loại sử dụng khối trích xuất đặc trưng tương tự như mô đun truy xuất ảnh thực vật. Tuy nhiên, mô hình phân loại còn kết hợp thêm một tầng mạng neuron đa lớp sau khi trích xuất đặc trưng, đồng thời huấn luyện sử dụng hàm mất mát cross entropy thông thường để phục vụ cho việc phân loại. Toàn bộ quy trình bao gồm các bước sau:

- **Tiền xử lý:** ảnh đầu vào được resize về kích thước 224x224 pixels, chuẩn hóa giá trị. Riêng với quá trình huấn luyện, các bước tăng cường dữ liệu (xoay ảnh, đổi màu, crop ảnh) được áp dụng để tăng tính tổng quát hóa của mô hình.

- **Trích xuất đặc trưng:** ảnh sau tiền xử lý sẽ được đưa vào khối Dual-Stream Fusion. Các đặc trưng sau khi được trích xuất từ đây sẽ được đưa vào đầu phân loại (Multi-layer perceptron với 2 layers + ReLU + Softmax).

**- Tổng hợp kết quả (nếu có nhiều ảnh):**

- + Từng ảnh được phân loại riêng biệt để thu được vector độ tin cậy riêng.
- + Các vector được cộng trọng số theo độ tin cậy.
- + Dự đoán và lựa ra top-k cuối cùng dựa trên vector tổng hợp.

Chiến lược này giúp cải thiện độ chính xác khi có nhiều ảnh của cùng một cây.

### 3.4. Truy xuất văn bản

Mô đun truy xuất văn bản có nhiệm vụ truy xuất ngữ cảnh thực vật phù hợp từ cơ sở dữ liệu tri thức, nhằm hỗ trợ mô đun sinh văn bản tạo câu trả lời chính xác. Các chức năng chính của mô đun này bao gồm:

- **Mã hóa văn bản ngữ nghĩa (Semantic Text Embedding):** mục tiêu chính của đồ án tập trung vào việc xây dựng cơ sở dữ liệu tri thức cho thực vật, vì vậy, tôi quyết định khảo sát các giải pháp hiện có và áp dụng vào bài toán truy xuất văn bản trong lĩnh vực thực vật, thay vì nghiên cứu một giải pháp mới cho vấn đề này. Dựa trên bảng xếp hạng Massive Text Embedding Benchmark [49], BGE-M3 cho thấy khả năng cân bằng giữa chất lượng kết quả và hiệu suất – khi là một trong những mô hình mạnh mẽ nhất với kích thước dưới 1 tỷ tham số. Đồng thời, khả năng của BGE-M3 không chỉ dừng lại ở kết quả tốt trong cùng phân khúc kích thước, mà còn đến từ khả năng hỗ trợ đa ngôn ngữ, bao quát cả tiếng Việt.

- **Mã hóa văn bản từ khóa:** Bên cạnh phương pháp mã hóa ngữ nghĩa bằng vector dày đặc, mô-đun truy xuất còn tích hợp mã hóa thưa dựa trên từ khóa, sử dụng thuật toán BM25 – một trong những phương pháp truy xuất từ khóa hiệu quả hiện nay. Việc lựa chọn BM25 thay vì các phương pháp truyền thống như TF-IDF xuất phát từ khả năng khắc phục nhiều điểm yếu có hữu trong xử lý văn bản ngắn và trung bình. Khác với TF-IDF vốn tuyến tính theo tần suất từ, BM25 áp dụng hàm bão hòa phi tuyến cho tần suất, giúp hạn chế việc những từ xuất hiện quá thường xuyên chi phối điểm số không tương xứng. Bên cạnh đó, BM25 còn tích hợp yếu tố điều chỉnh độ dài tài liệu, đảm bảo sự công bằng khi so sánh giữa các văn bản có độ dài khác biệt – một yếu tố đặc biệt quan trọng trong miền dữ liệu y học, nơi các mô tả về công dụng, thành phần hay tương tác thuốc thường có cấu trúc và độ dài không đồng nhất.

- **Tìm kiếm tương đồng (Similarity Search):** Khác với mô đun truy xuất ảnh sử dụng phương pháp tìm kiếm đơn lẻ, mô đun truy xuất văn bản áp dụng chiến lược tìm kiếm kết hợp (hybrid search) để tối ưu hóa hiệu quả truy xuất. Hệ thống thực hiện song song hai quá trình tìm kiếm độc lập: tìm kiếm ngữ nghĩa thông qua BGE-M3 để nắm bắt ý nghĩa sâu của câu hỏi, và tìm kiếm từ khóa thông qua BM25 để đảm bảo không bỏ sót các thuật ngữ chuyên môn quan trọng. Kết quả từ hai kênh tìm kiếm sau đó được tổng hợp bằng kỹ thuật Reciprocal Rank Fusion (RRF), một phương pháp đã được chứng minh hiệu quả trong việc kết hợp các danh sách xếp hạng khác nhau. RRF tính toán điểm số tổng hợp cho mỗi tài liệu dựa trên vị trí thứ hạng của nó trong các danh sách kết quả, ưu tiên những tài liệu xuất hiện ở vị trí cao trong nhiều kênh tìm kiếm. Cách tiếp cận này giúp cân bằng giữa khả năng hiểu ngữ cảnh của dense retrieval và độ chính xác từ khóa của sparse retrieval, từ đó cung cấp kết quả truy xuất toàn diện và chính xác hơn cho các câu hỏi về y học cổ truyền và thực vật.

Đầu vào và đầu ra của mô đun này có thể được mô tả như sau:

- **Input:** câu hỏi của người dùng
- **Output:** danh sách bao gồm các văn bản ngữ cảnh liên quan đến câu hỏi, được sắp xếp theo mức độ liên quan

Sự thành công của mô đun này phụ thuộc rất nhiều vào chất lượng và độ liên quan giữa cơ sở dữ liệu tri thức mà nó truy xuất với câu hỏi của người dùng và lĩnh vực áp dụng. Sau khi phân tích metadata thực vật đã thu thập được, tôi nhận thấy có rất nhiều loại thực vật có ứng dụng trong y học cổ truyền. Dựa trên phát hiện này, tôi đã tập trung xây dựng một cơ sở tri thức chuyên biệt về ứng dụng y học của thực vật rừng khu vực Đà Nẵng. Để làm được điều này, tôi xây dựng một pipeline gồm 2 giai đoạn:

- Tăng cường metadata đã thu thập.
- Trích xuất thông tin về ứng dụng y học từ metadata trên.

### **Tăng cường metadata đã thu thập**

Để cải thiện độ phong phú của dữ liệu, bên cạnh các dữ liệu đã có, tôi thu thập thêm dữ liệu từ sách “Những cây thuốc và vị thuốc Việt Nam” của Đỗ Tất Lợi – một tài liệu rất có giá trị về dược liệu Việt Nam. [51]

- Để trích xuất thông tin từ sách, tôi tận dụng Gemini của Google [52] – một mô hình đa thê thức mạnh mẽ, để thực hiện nhận diện ký tự quang học (OCR) trên các trang sách.
- Sau khi chuyển toàn bộ nội dung sách thành văn bản, tôi tiếp tục sử dụng Gemini với các prompt có cấu trúc để trích xuất thông tin chính xác ở cấp độ loài.

- Kết quả, tôi đã thu thập được dữ liệu chi tiết cho hơn 600 loài thực vật. Tuy nhiên, vì đề tài tập trung vào khu vực Đà Nẵng, tôi tiến hành đối chiếu chéo với metadata hiện có để loại bỏ các loài không thuộc khu vực này. Các bản ghi còn lại được hợp nhất vào metadata ban đầu, giúp nâng cao đáng kể độ phong phú của dữ liệu.

### **Trích xuất thông tin về ứng dụng y học**

Trích xuất tri thức về các ứng dụng y học của thực vật là trọng tâm trong xây dựng cơ sở tri thức này. Thay vì dựa vào metadata thuần hiện có, tôi phát triển một pipeline trích xuất nhiều vòng (multi-pass extraction), kết hợp cùng sức mạnh của mô hình ngôn ngữ lớn (Gemini) để trích xuất dữ liệu tri thức, có cấu trúc, phong phú, từ metadata gốc. Pipeline gồm 4 vòng, tận dụng Gemini kết hợp với prompt tiếng Việt được thiết kế kỹ lưỡng. Mỗi vòng nhắm đến một lớp tri thức y học cụ thể.

### **Vòng 1: Trích xuất ứng dụng điều trị**

- Xác định mối quan hệ điều trị trực tiếp, gồm:
  - + Bệnh và triệu chứng mà mỗi cây thuốc được sử dụng để điều trị.
  - + Mức độ hiệu quả (cao, trung bình, thấp).
  - + Phương pháp bào chế, liều lượng truyền thống.
  - + Điểm tin cậy cho từng mối quan hệ điều trị.
- Hệ thống chuẩn hóa tên bệnh (ví dụ: đau dạ dày -> stomach\_pain) nhưng vẫn giữ song song thuật ngữ tiếng Việt gốc để đảm bảo rõ ràng và giữ được ngữ cảnh văn hóa.

### **Vòng 2: Mối quan hệ tương tác**

- Trong y học cổ truyền, có nhiều cây thuốc không được sử dụng đơn lẻ, vì vậy, vòng này trích xuất:
  - + Các loài được dùng chung trong các bài thuốc đa thành phần.
  - + Loại tương tác (tăng cường, hỗ trợ, phụ thuộc).
  - + Hoàn cảnh sử dụng điển hình của kết hợp đó.
  - + Tỷ lệ, liều lượng.

### **Vòng 3: Trích xuất các công thức y học cổ truyền**

- Trích xuất các bài thuốc truyền thống hoàn chỉnh, bao gồm:
  - + Tên bài thuốc (ví dụ: Tú ché hương phụ, Cao hương ngải).
  - + Danh sách thành phần kèm liều lượng cụ thể.
  - + Phương pháp bào chế, các yếu tố mùa vụ hoặc thời gian sử dụng.
  - + Hướng dẫn sử dụng, mục tiêu điều trị.

### **Vòng 4: Kiểm tra chất lượng và chuẩn hóa**

- Vòng này giúp đảm bảo tính nhất quán và độ tin cậy của dữ liệu đã trích xuất:
  - + Đổi chiều chéo giữa các thực thể và mối quan hệ đã trích xuất.

- + Phát hiện mâu thuẫn hoặc khoảng trống logic.
- + Gán điểm tin cậy tổng thể (thang 0-1) cho mỗi khối dữ liệu.
- + Đề xuất cải thiện hoặc xem xét dữ liệu thủ công nếu cần.

Sau pipeline 2 giai đoạn trên, tôi đã tạo ra một cơ sở tri thức phong phú về ứng dụng y học của thực vật rừng tại Đà Nẵng. Nhờ vào cơ sở tri thức này, mô đun truy xuất văn bản có khả năng cung cấp câu trả lời chính xác, giàu thông tin cho các câu hỏi liên quan đến cây thuốc.

Dựa trên loại câu hỏi mà người dùng cung cấp, mô đun sẽ quyết định xem nguồn dữ liệu để trả lời câu hỏi sẽ được lấy từ đâu bằng các quy tắc định sẵn (rule-based). Như đã trình bày ở mục 3.1, tôi chia các câu hỏi thành 3 loại:

- **Câu hỏi cụ thể về một loài thực vật:** đây là các câu hỏi đi kèm sau khi người dùng gửi cùng ảnh hoặc có cung cấp tên loài thực vật trong câu hỏi. Đối với loại câu hỏi này, thông tin bổ sung sẽ được lấy từ dữ liệu tổng hợp ban đầu để hỗ trợ quá trình sinh câu trả lời.

- **Câu hỏi về y học, tác dụng được lý, được tính của các loài thực vật:** với các câu hỏi thuộc loại này, ngữ cảnh cần thiết để trả lời sẽ được truy xuất từ cơ sở dữ liệu tri thức được xây dựng như đã nêu trên. Quá trình truy xuất gồm 2 bước sau:

+ **Truy xuất:** sử dụng truy xuất dày đặc để tìm kiếm các thông tin liên quan nhất về mặt ngữ nghĩa toàn cục bằng so sánh tương đồng vector thông qua FAISS và BGE-M3. Đồng thời, sử dụng truy xuất thưa với các vector biểu diễn tạo thành từ BM25 để tìm kiếm các ngữ cảnh phù hợp và chứa các từ khóa quan trọng.

+ **Tổng hợp kết quả:** sử dụng thuật toán RRF như đã trình bày ở 2.4.3 để tổng hợp danh sách kết quả cuối cùng. Sau đó sẽ lựa ra top-k văn bản từ danh sách này làm ngữ cảnh cho câu hỏi.

- **Câu hỏi chung:** các câu hỏi không thuộc 2 loại trên. Các câu hỏi này sẽ được xử lý bằng công cụ web-search có sẵn của Gemini API.

### 3.5. Sinh văn bản

Mô đun sinh văn bản (hay sinh câu trả lời) đóng vai trò là giao diện ngôn ngữ tự nhiên của hệ thống, chuyển hóa tri thức thực vật có cấu trúc thành các câu trả lời tiếng Việt dễ hiểu cho người dùng. Dù cho mô đun này không phải là trọng tâm nghiên cứu của đề tài, nó vẫn đóng vai trò quan trọng trong việc giúp người dùng dễ dàng tiếp cận và khai thác các chức năng nhận diện thực vật và truy xuất thông tin một cách tự nhiên. Các chức năng chính của mô đun bao gồm:

- **Sinh câu trả lời có hỗ trợ RAG:** kết hợp chặt chẽ với mô đun truy xuất văn bản, khai thác từ cơ sở tri thức y học đã trích xuất để sinh các câu trả lời giàu thông tin.

- **Khai thác từ dữ liệu web (Web-augment Knowledge Access):** đối với những câu hỏi chung chung, không có ngữ cảnh trong cơ sở dữ liệu, mô đun có khả năng tận dụng tìm kiếm web để cung cấp câu trả lời chính xác hơn.

Đầu vào và đầu ra của mô đun được mô tả như sau:

- **Input:** Câu hỏi của người dùng + tên loài thực vật (nếu có) + thông tin liên quan đến câu hỏi (nếu có)

- **Output:** Câu trả lời hoàn chỉnh, phù hợp cho câu hỏi

Đối với mô đun này, một quyết định then chốt là lựa chọn phương pháp triển khai mô hình ngôn ngữ lớn (LLM). Như đã phân tích tại mục 2.4.6, hai hướng tiếp cận chính là tự triển khai (self-hosted) và sử dụng dịch vụ qua API (API-hosted). Xét trong bối cảnh của đề tài, với trọng tâm là xây dựng hệ thống nhận diện và truy xuất thông tin thực vật thay vì phát triển LLM, phương án sử dụng API-hosted được lựa chọn vì những lý do thực tiễn và chiến lược sau:

- **Hạn chế tài nguyên:** việc xây dựng và triển khai LLM cục bộ đòi hỏi tài nguyên tính toán (GPU) đáng kể, thông thường, để triển khai một LLM với kích thước 7 tỉ tham số - một kích thước trung bình để có thể cho ra chất lượng chấp nhận được với tiếng Việt sẽ đòi hỏi hơn 16GB. Đây là một kích thước mà chỉ có các GPU cao cấp với giá thành đắt đỏ mới có thể đáp ứng. Đối với một đề tài mà trọng tâm là nhận diện thực vật và truy xuất thông tin, đầu tư quá nhiều vào tài nguyên phục vụ LLM sẽ ảnh hưởng đến phần cốt lõi.

- **Chất lượng tiếng Việt:** các mô hình mã nguồn mở hiện tại (Qwen, Gemma, LLaMa, Mistral, ...) có chất lượng thấp hơn đáng kể so với Gemini, chưa kể đến chất

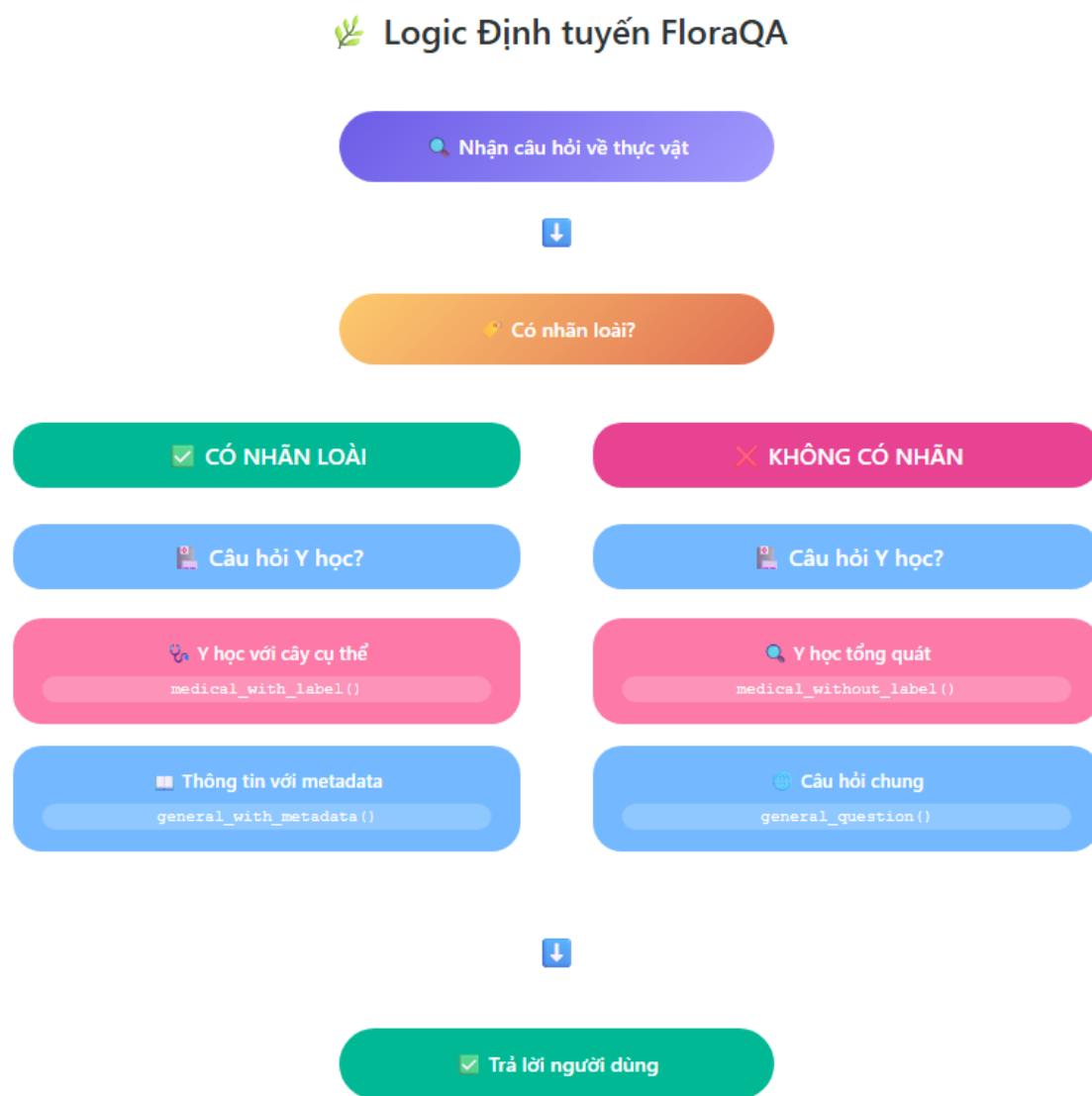
lượng tiếng Việt. Với một lĩnh vực cần truyền đạt đầy đủ và chính xác thông tin như y học cổ truyền và thuật ngữ thực vật học, việc sinh ra các thông tin sai lệch là điều tối kị. Qua các thử nghiệm ban đầu, các mô hình cục bộ cho các kết quả tiếng Việt không ổn định như Gemini.

- **Tốc độ phát triển và ưu tiên nghiên cứu:** thiết lập và tối ưu hạ tầng phục vụ LLM cục bộ đòi hỏi nhiều công sức kỹ thuật, không phù hợp với mục tiêu chính của đề tài. API-based giúp tối tập trung vào các nghiên cứu và mục tiêu chính.

- **Chi phí và khả năng vận hành, mở rộng:** Với quy mô thử nghiệm của đề tài hiện tại, phương án sử dụng mô hình LLM qua API (trả phí theo mức sử dụng) tỏ ra kinh tế và linh hoạt hơn nhiều so với việc đầu tư tài nguyên phần cứng để triển khai mô hình cục bộ. Đặc biệt, khi nhu cầu mở rộng trong tương lai phát sinh, phương án API cho phép mở rộng gần như tức thời mà không cần thay đổi hạ tầng hay mã nguồn. Ngược lại, nếu sử dụng LLM cục bộ, việc nâng cấp khả năng tính toán hoặc mở rộng kích thước mô hình sẽ phát sinh chi phí phần cứng đáng kể và đòi hỏi điều chỉnh phức tạp trong hệ thống.

Từ những phân tích trên, có thể thấy rằng việc tự triển khai một LLM self-hosted sẽ tạo ra nhiều rào cản về tài nguyên, chất lượng và tiến độ, không phù hợp với mục tiêu của đề tài. Do đó, quyết định chiến lược là tận dụng dịch vụ API LLM hiện đại (cụ thể là Gemini của Google) để đảm bảo hệ thống có khả năng sinh văn bản chất lượng cao một cách hiệu quả và linh hoạt.

Trong mô-đun này, câu hỏi cùng ngữ cảnh liên quan (được trích xuất từ các bước xử lý trước đó) sẽ được gửi đến mô hình Gemini thông qua API. Thay vì chỉ đơn thuần chuyển tiếp câu hỏi và ngữ cảnh, hệ thống đã được thiết kế để bổ sung các hướng dẫn chi tiết về cách trả lời, kiểm chứng và xác thực thông tin, nhằm nâng cao độ chính xác và mức độ tin cậy của phản hồi. Kết quả sau cùng được hiển thị trực tiếp lên giao diện người dùng.



Hình 3-2: Luồng xử lý câu hỏi của FloraQA

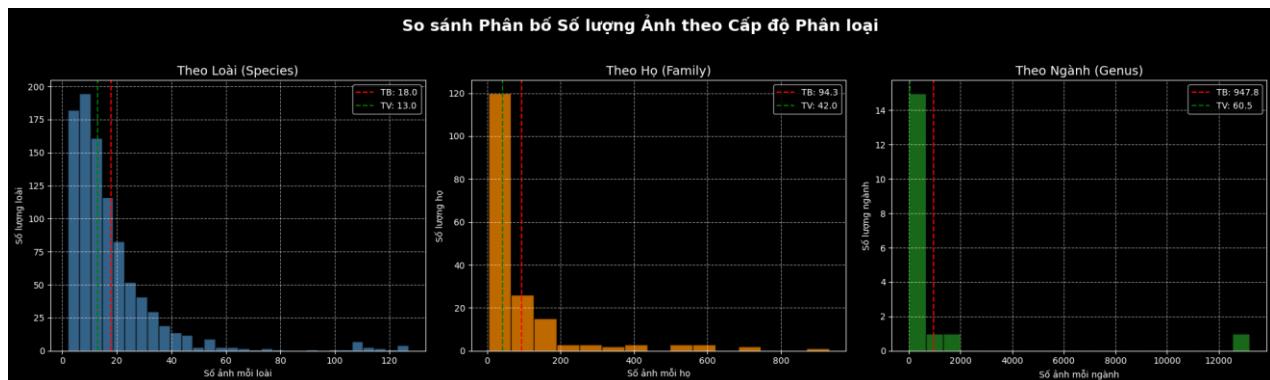
## CHƯƠNG 4: TRIỂN KHAI VÀ ĐÁNH GIÁ KẾT QUẢ

### 4.1. Thực nghiệm huấn luyện

#### 4.1.1. Bộ dữ liệu

Bộ dữ liệu được sử dụng trong đề tài bao gồm hơn 17,000 ảnh, đại diện cho 949 loài thực vật đã biết có mặt trong các khu rừng của Đà Nẵng. Các ảnh này được thu thập từ nhiều nguồn khác nhau, bao gồm: ảnh chụp thủ công, kết quả tìm kiếm hình ảnh từ Google, các trang thông tin trên Internet.

Mặc dù số lượng mẫu lớn, tập dữ liệu vẫn có sự thiên lệch, chủ yếu là ảnh của các lá cây trong trạng thái khỏe mạnh, không bị tác động - chưa đa dạng về các trường hợp bệnh lý, lão hóa, thương tổn vật lý hay ảnh của các bộ phận khác như hoa, quả. Điều này tạo ra một thách thức đáng kể cho mô hình, do hạn chế khả năng khai quát hóa đối với các biến thiên trong loài khi triển khai trong môi trường thực tế.



Hình 4-1: Biểu đồ phân phối số lượng ảnh theo loài - họ - ngành trong tập dữ liệu

Biểu đồ phân bố dữ liệu cho thấy một thực trạng rõ rệt: số lượng ảnh ở các cấp độ phân loại từ loài, họ đến ngành đều thể hiện sự mất cân bằng nghiêm trọng, với đặc trưng là phân bố lệch phải (right-skewed). Mức độ mất cân bằng này trở nên trầm trọng hơn ở các cấp độ phân loại cao hơn như họ và đặc biệt là ngành. Tình trạng này đặt ra một thách thức lớn, bởi một mô hình được huấn luyện trên dữ liệu như vậy sẽ có xu hướng thiên vị (bias), học chủ yếu từ các lớp đa số (ngành/họ có nhiều ảnh) và bỏ qua các lớp thiểu số.

Cũng bởi phát hiện này đã giúp tôi quyết định: thay vì áp dụng phương pháp phân loại thứ bậc truyền thống, mô hình sẽ thực hiện phân loại trực tiếp ở cấp độ loài (thể hiện

qua việc mô đun phân loại thực vật chỉ sử dụng 1 head phân loại cơ bản gồm 1 tầng MLP 2 lớp). Lựa chọn này dựa trên cơ sở cấp loài là cấp có độ lệch thấp nhất, giúp giảm thiểu tác động tiêu cực của sự mất cân bằng. Nguyên nhân của sự chênh lệch đến từ đặc thù đa dạng sinh học của khu vực thu thập: dù phong phú, phần lớn các loài ở Đà Nẵng lại tập trung trong một số họ và ngành nhất định. Qua đó, có thể khẳng định vai trò thiết yếu của việc phân tích dữ liệu thăm dò (Exploratory Data Analysis - EDA), giúp lựa chọn phương pháp luận phù hợp thay vì áp dụng máy móc các cách tiếp cận kinh điển.

#### **4.1.2 Cấu hình huấn luyện**

Trong phạm vi đề tài này, tôi chỉ thực hiện huấn luyện mô hình Dual-Stream cho các tác vụ Nhận diện thực vật và Truy xuất ảnh thực vật. Đối với các tác vụ Truy xuất văn bản và Sinh câu trả lời, như đã trình bày, tôi sử dụng trực tiếp mô hình BGE-M3 và API Gemini, không thực hiện tinh chỉnh thêm.

Đối với hai tác vụ nhận diện và truy xuất ảnh thực vật, tôi sử dụng phiên bản pre-trained của Vision Transformers và phiên bản pre-trained của ConvNeXtV2 tạo thành mô hình Dual-Stream Fusion.

Việc huấn luyện cả 2 mô hình được thực hiện trên cùng cấu hình và cài đặt như sau

Bảng 4-1: Cấu hình huấn luyện các mô hình xử lý ảnh

CPU	Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz
GPU	NVIDIA RTX 3090
Epoch	100
Batch size	64

Cả hai mô hình xử lý ảnh đều được huấn luyện với cùng tham số, cấu hình. Điểm khác biệt duy nhất nằm ở hàm mất mát. Trong khi mô hình phân loại sử dụng hàm mất mát Cross Entropy thông thường, thì mô hình mã hóa ảnh sử dụng hàm mất mát ProxyNCA loss như đã trình bày ở mục 2.2.2.

## 4.2. Đánh giá kết quả

### 4.2.1. Đánh giá kết quả mô hình phân loại

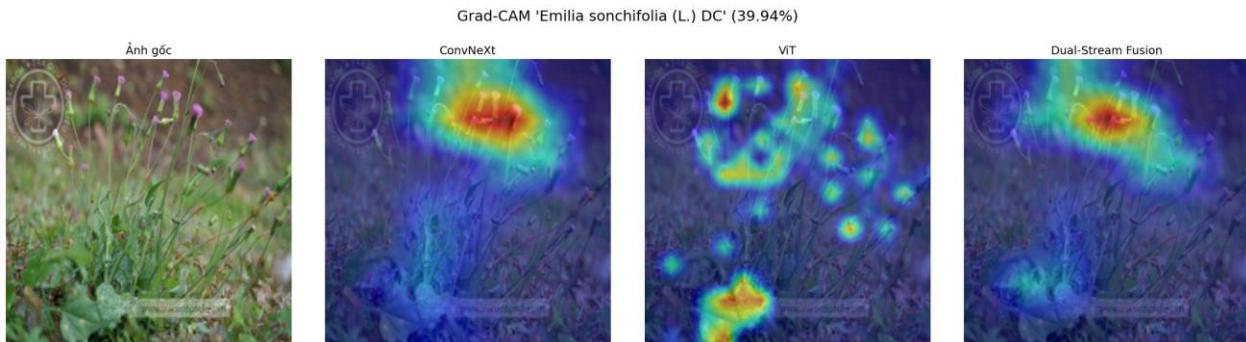
Trước hết, tôi đánh giá kết quả của mô hình phân loại trên tập kiểm thử. Tôi thực hiện việc so sánh với cả Vision Transformers và ConvNeXt V2. Kết quả được trình bày ở bảng sau:

Bảng 4-2: Bảng đánh giá kết quả mô hình phân loại

	Accuracy@1	Accuracy@5	Recall	Precision	F1-score
Vision Transformers	0.6941	0.8442	0.6941	0.6576	0.6531
ConvNeXtV2	0.6891	0.8383	0.6891	0.6579	0.6471
<b>Dual-Stream Fusion</b>	<b>0.7583</b>	<b>0.8705</b>	<b>0.7583</b>	<b>0.7730</b>	<b>0.7452</b>

Bảng kết quả cho thấy sự vượt trội rõ rệt của kiến trúc đề xuất Dual-Stream Fusion so với các kiến trúc tiên tiến khác trên toàn bộ các tiêu chí đánh giá. Đáng chú ý, với Accuracy@5 đạt **0.8705**, điều này đã thành công hoàn thành mục tiêu đầu tiên trong các mục tiêu của đề tài này.

Để phân tích sâu hơn tại sao kiến trúc này lại hiệu quả, tôi sử dụng Grad-CAM [53] để trực quan hóa vùng quan tâm của mô hình. Hình 4.2 cho thấy một ví dụ điển hình cho quá trình xử lý của Dual-Stream Fusion.



Hình 4-2: Feature map từ các mô hình khác nhau

- **Nhánh ConvNeXt**, với thế mạnh về xử lý đặc trưng cục bộ, đã tập trung sự chú ý vào vùng có mật độ thông tin cao nhất – cụm hoa dày đặc ở trung tâm ảnh. Vùng kích hoạt (heatmap) lớn và liền mạch, thể hiện khả năng nhận diện các mẫu hình và kết cấu (texture) tại một khu vực cụ thể.

- **Nhánh Vision Transformer (ViT)**, ngược lại, thể hiện khả năng nắm bắt ngữ cảnh toàn cục nhờ cơ chế self-attention. Các vùng kích hoạt của nó phân tán trên toàn bộ ảnh, xác định chính xác vị trí của từng bông hoa riêng lẻ và cả một chiếc lá đặc trưng ở phía dưới.
- **Kết quả của Dual-Stream Fusion** là một sự kết hợp thông minh. Mô hình đã kế thừa sự tập trung mạnh mẽ vào cụm hoa chính từ ConvNeXt, nhưng đồng thời vẫn duy trì "nhận thức" về chiếc lá quan trọng mà ViT đã chỉ ra. Điều này chứng tỏ kiến trúc đề xuất đã học được cách cân bằng hiệu quả giữa việc phân tích chi tiết cục bộ và bối cảnh tổng thể, tạo ra một biểu diễn đặc trưng toàn diện hơn để đưa ra quyết định cuối cùng

Những kết quả này khẳng định rằng kiến trúc Dual-Stream Fusion đã tạo ra một sự cân bằng xuất sắc giữa hiệu suất và hiệu quả tính toán. Điều này chứng tỏ tiềm năng to lớn của kiến trúc trong các ứng dụng thực tế, nơi hiệu quả về tài nguyên là một yếu tố then chốt, cho phép đạt được độ chính xác hàng đầu mà không cần đến các mô hình khổng lồ.

#### **4.2.2 Đánh giá kết quả mô hình truy xuất ảnh**

Mục đích của mô hình truy xuất ảnh ở trong hệ thống là bộ lọc các ảnh trong và ngoài miền, vì vậy, khác với những cách đánh giá truy xuất ảnh thuần túy (đánh giá các ảnh truy xuất ra có tương tự với ảnh đầu vào hay không), tôi thực hiện xây dựng một bộ dữ liệu kiểm thử chỉ gồm 2 lớp: trong miền (in-domain) và ngoài miền (out-domain). Bộ dữ liệu cho các ảnh trong miền được lấy từ bộ dữ liệu kiểm thử của mô hình phân loại, trong khi bộ dữ liệu cho lớp ngoài miền bao gồm hơn 2000 ảnh được tôi thu thập từ các nguồn trên Internet, với các ảnh thuộc 3 chủ đề chính:

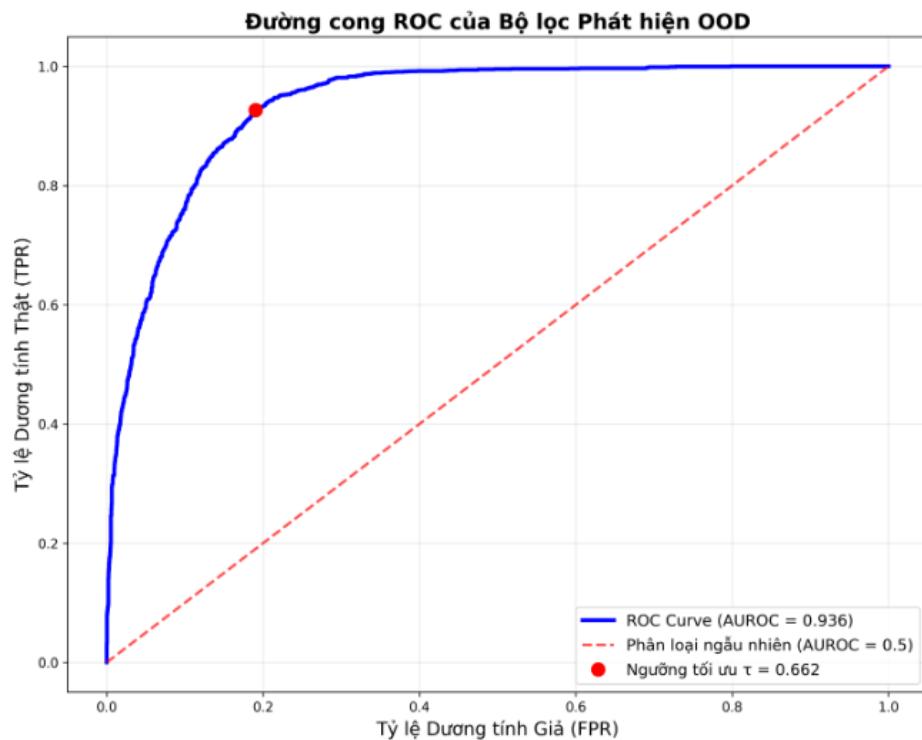
- **Đồ vật:** các vật dụng bình thường
- **Động vật:** hình ảnh các loài động vật
- **Thực vật:** hình ảnh của các loài cây khác, không nằm trong tập dữ liệu

Từ trên bộ dữ liệu này, đánh giá được thực hiện bằng cách áp dụng phương pháp trung bình lớp như đã đề cập ở mục 2.3.2 để xác định các ảnh trong/ngoài miền. Kết quả được trình bày trong bảng sau:

Bảng 4-3: Bảng đánh giá kết quả mô hình truy xuất ảnh

	AUROC	TNR (Loại bỏ OOD)	TPR (Chấp nhận ID)	Accuracy	F1-Score	FPR@95TPR
Vision Transformer	0.9272	<b>0.8122</b>	0.8998	0.8460	0.8193	0.2555
ConvNeXt V2	0.9125	0.7929	0.8850	0.8284	0.8001	0.2915
<b>Dual-Stream Fusion</b>	<b>0.9359</b>	0.8096	<b>0.9266</b>	<b>0.8548</b>	<b>0.8320</b>	<b>0.2206</b>

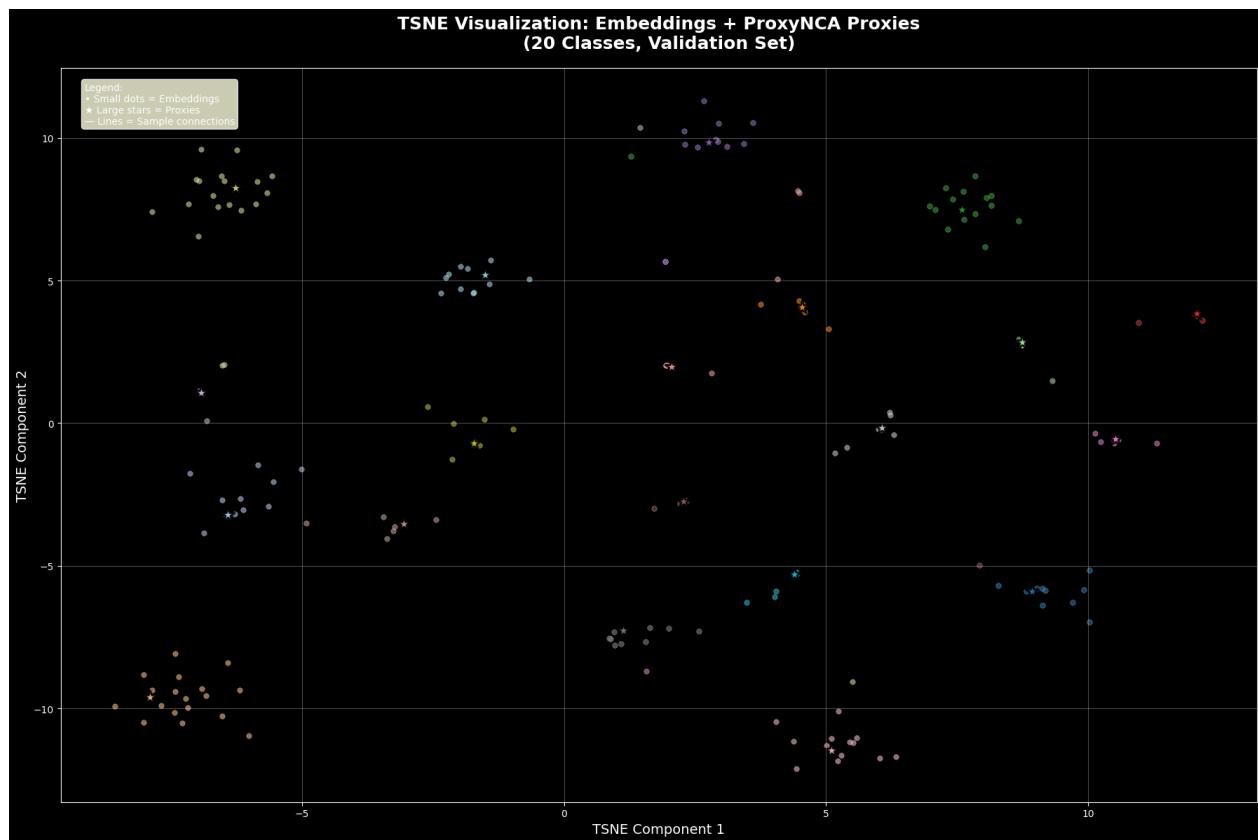
Từ bảng trên, quan sát thấy rằng Dual-Stream Fusion thể hiện hiệu suất vượt trội, với chỉ số AUROC đạt **0.9359**, cao nhất trong ba mô hình. Chỉ số AUROC càng cao cho thấy mô hình có khả năng phân biệt tốt giữa ảnh ID và OOD trên toàn bộ dải ngưỡng.



Hình 4-3: Đường cong ROC của mô hình truy xuất ảnh với backbone Dual-Stream Fusion

Đặc biệt, tại mức TPR = 95%, Dual-Stream Fusion đạt tỷ lệ dương tính giả (FPR) thấp nhất là **22.06%**, khẳng định năng lực của mô hình trong việc giữ an toàn cho hệ thống ngay cả khi ưu tiên tối đa việc chấp nhận các ảnh hợp lệ.

Để quan sát kĩ hơn tác động của mô đun này, tôi sử dụng t-SNE [59] để trực quan hóa không gian vector của 20 loài thực vật ngẫu nhiên sau khi đã giảm chiều xuống 2D. Có thể thấy, các loài khác nhau được phân bố ở các vùng nhất định, giải thích cho kết quả tốt của mô hình trong tác vụ truy xuất ảnh.



Hình 4-4: Minh họa phân bố của 20 lớp thực vật ngẫu nhiên tạo bởi mô hình truy xuất ảnh

#### **4.2.3. Đánh giá kết quả mô hình truy xuất văn bản**

Để đánh giá được hiệu suất của mô đun truy xuất văn bản, tôi tận dụng Gemini API để xây dựng một bộ dữ liệu kiểm thử, gồm 75 truy vấn đi kèm tài liệu liên quan, dựa trên dữ liệu y học thu thập được qua quá trình đề cập ở mục 3.4. Đánh giá được thực hiện trên 3 phương pháp truy xuất văn bản: Dense Retrieval, Sparse Retrieval và Hybrid Retrieval. Kết quả được trình bày ở bảng sau:

Bảng 4-5: Bảng đánh giá kết quả mô hình truy xuất văn bản

	Hit Rate	MRR	NDCG
Dense	0.7162	0.6234	0.6447
Sparse	0.6622	0.4849	0.5225
<b>Hybrid</b>	<b>0.7838</b>	<b>0.6768</b>	<b>0.6938</b>

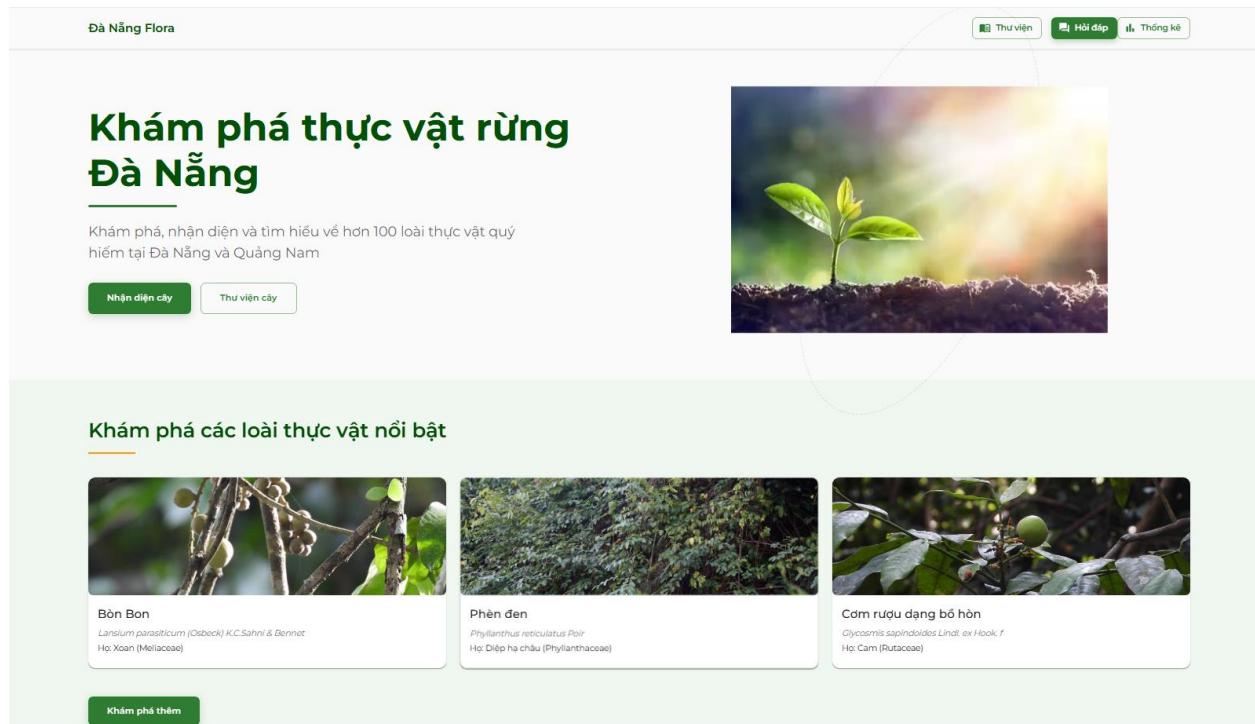
Từ bảng trên có thể thấy, cách tiếp cận Hybrid quả thực vượt trội hơn so với từng cách tiếp cận đơn lẻ. Cụ thể hơn, Sparse Retrieval cho kết quả tốt với những câu truy vấn đề cập đến các khái niệm cụ thể như “Cây thuốc nào hiệu quả để điều trị tiêu chảy?”. Trong khi đó, Dense Retrieval lại không tỏ ra vượt trội ở những truy vấn đơn giản này – nguyên nhân nằm ở nội dung của các văn bản có sự chồng chéo lớn, khi có rất nhiều các loài thực vật có cùng tác dụng chữa các căn bệnh thường gặp như: tiêu chảy, ho, sốt. Với sự trùng lặp về mặt nội dung lớn như vậy, Dense Retrieval tìm ra cơ sở các loài cây, nhưng bởi vì số lượng lớn khiến cho các loài cây ở ground truth không nằm trong top-k kết quả tìm ra được. Ngược lại, với những truy vấn phức tạp hơn như “Cây nào có thể chữa cả đau bụng và ăn không tiêu?”, Dense Retrieval lại tỏ ra hiệu quả hơn, với cùng lý do. Khi truy vấn chứa nhiều cụm từ quan trọng (đau bụng, ăn không tiêu), việc thực hiện keyword matching như Sparse Retrieval lúc này lại cho ra quá nhiều kết quả, và đa số kết quả có thể chỉ cần chứa 1 trong 2 căn bệnh trên, dẫn đến kết quả không cao. Trong khi Dense Retrieval có thể nắm bắt được ý nghĩa rằng cần phải tồn tại cả 2 căn bệnh này trong mục tác dụng của 1 loài cây, dẫn đến việc sàng lọc bỏ những loài thực vật chỉ có tác dụng với 1 trong 2 loại bệnh. Từ đó, việc kết hợp 2 phương pháp truy xuất trên bằng Hybrid Retrieval có thể giúp cho hệ thống đảm bảo được hiệu suất đối với những truy vấn từ đơn giản đến phức tạp hơn.

#### 4.3. Triển khai

Để mở rộng tính ứng dụng thực tiễn của đề tài này, tôi đã phát triển và triển khai hệ thống này lên nền tảng Internet, với mục đích đưa hệ thống đến gần hơn với người dùng. Điều này giúp cho FloraQA có thể trở nên phổ biến và được sử dụng rộng rãi bởi không chỉ các chuyên gia thực vật, sinh viên nghiên cứu mà còn cả số đông đại chúng.

#### 4.3.1. Giao diện ứng dụng

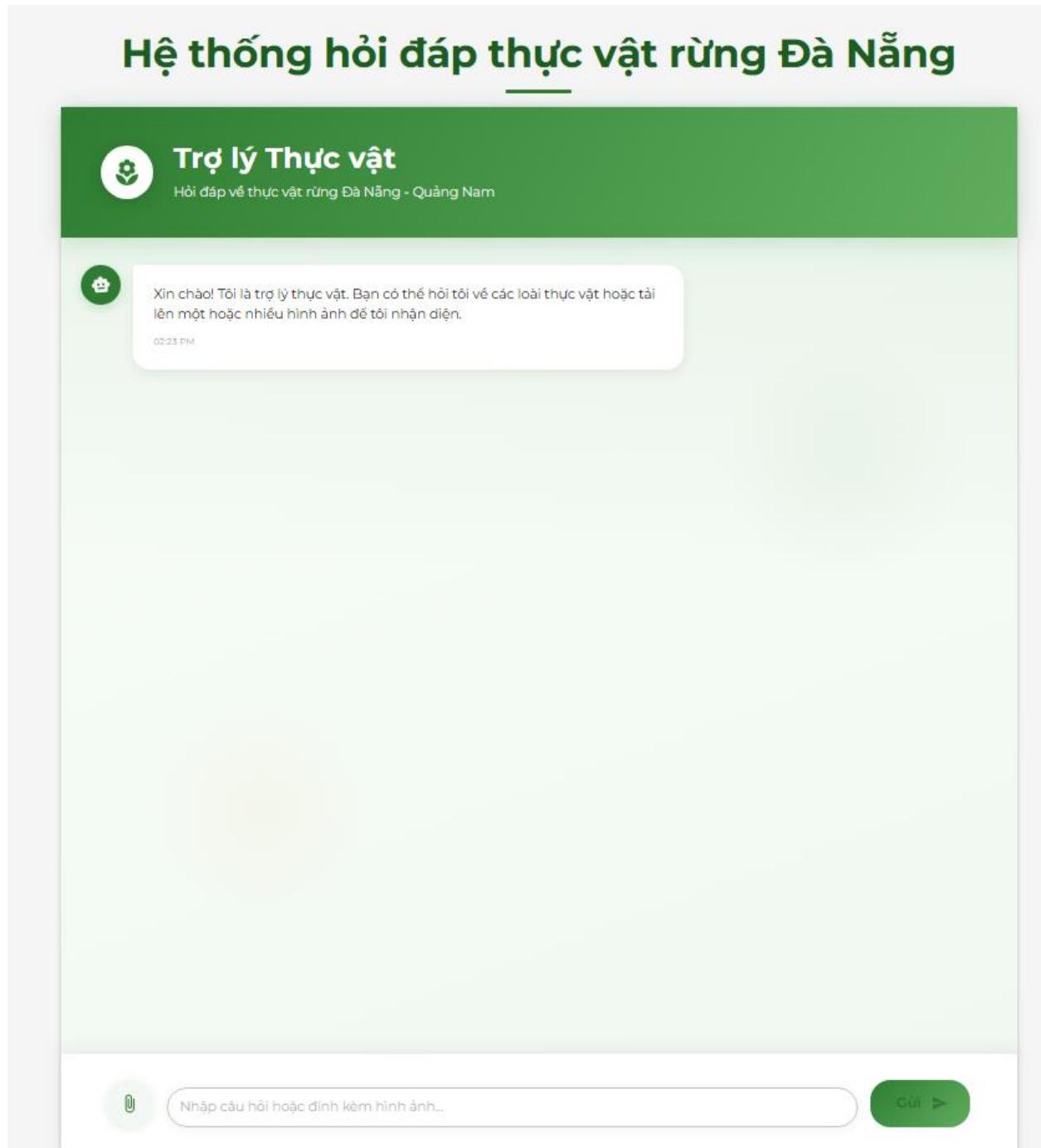
Từ trang chủ, người dùng có thể dễ dàng truy cập đến 2 tính năng lớn của trang web: hỏi đáp và thư viện. Tính năng thư viện trực quan hóa toàn bộ thông tin của các loài thực vật đã thu thập được, trong khi tính năng hỏi đáp – tính năng chính của hệ thống, sẽ có luồng hoạt động như đã mô tả ở mục 3.1.



Hình 4-5: Trang chủ

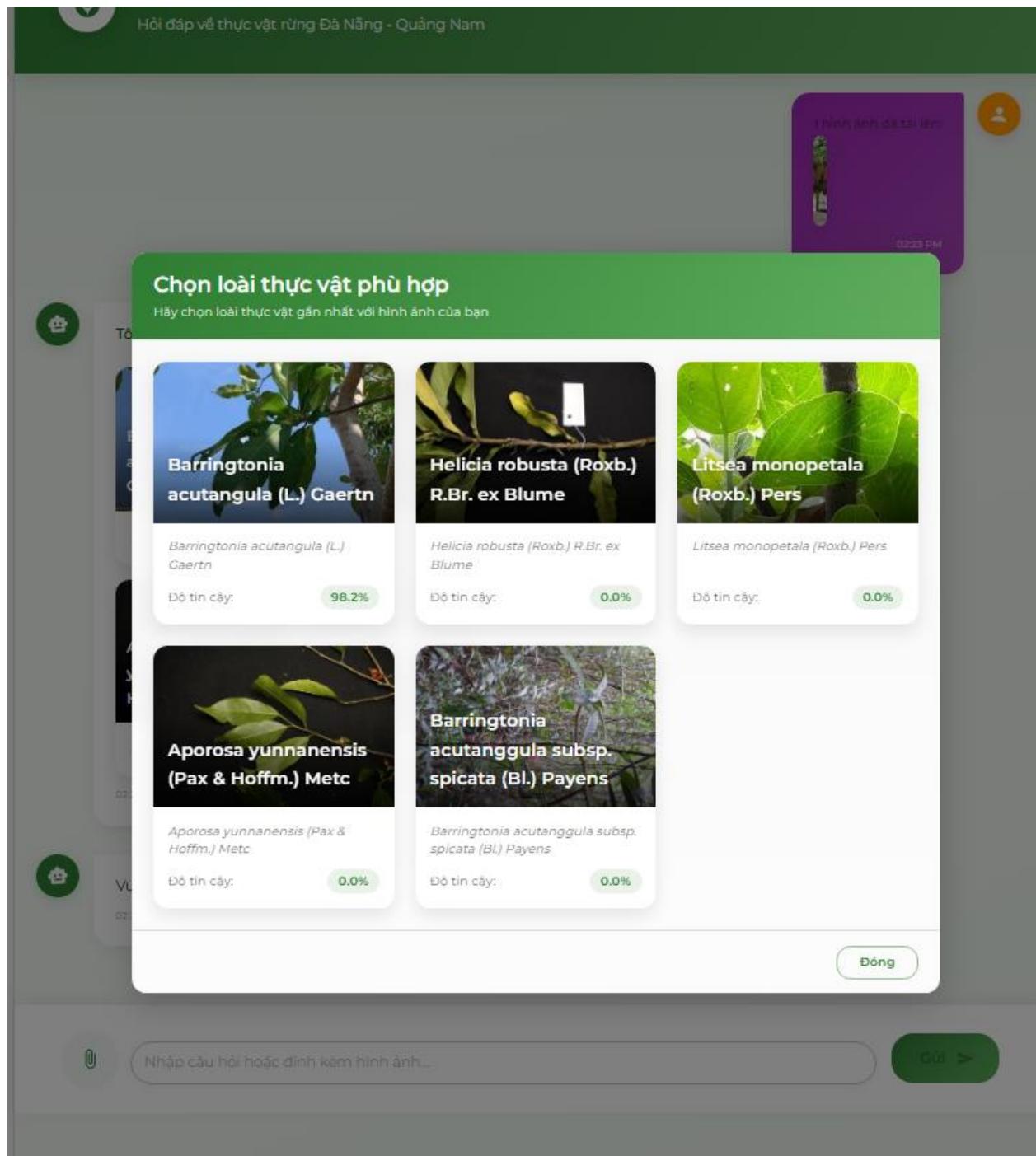
#### 4.3.2. Hỏi đáp

Lấy cảm hứng từ nhiều trang web cung cấp dịch vụ hỏi đáp như Gemini, ChatGPT, tôi xây dựng một giao diện đơn giản, thân thiện với người dùng. Khi vào tính năng hỏi đáp, người dùng có thể gửi câu hỏi, hình ảnh hoặc cả hai cùng lúc để kích hoạt các luồng khác nhau của hệ thống. Một số minh họa của tính năng được thể hiện qua các hình ảnh bên dưới



Hình 4-6 Giao diện tính năng hỏi đáp

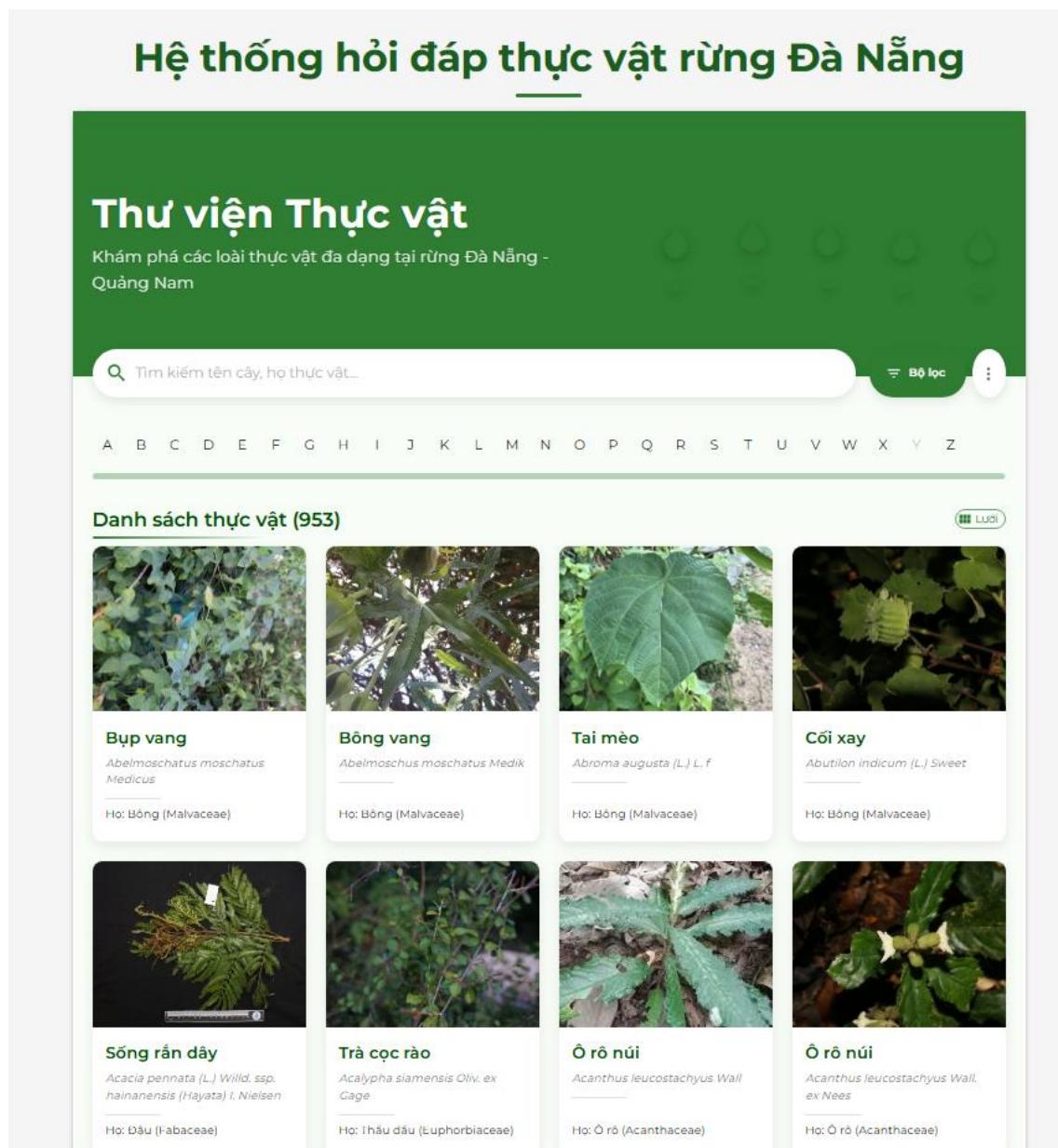
## Hệ thống hỏi đáp thực vật rừng khu vực Đà Nẵng



Hình 4-7: Giao diện tính năng hỏi đáp - 2

### 4.3.3. Thư viện

Bên cạnh tính năng chính là hỏi đáp về thực vật, hệ thống còn cung cấp một tính năng phụ là thư viện – nơi thể hiện toàn bộ các thông tin về các loài thực vật đã thu thập được, giúp người dùng có thể nhanh chóng tra cứu khi đã biết tên hoặc muốn tìm hiểu về một loài thực vật cụ thể nào đó. Một số giao diện của tính năng được thể hiện qua các hình ảnh bên dưới



Hình 4-8: Giao diện tính năng thư viện

## KẾT LUẬN

Qua quá trình nghiên cứu, khám phá lý thuyết và phát triển hệ thống, đề tài “Hệ thống Hỏi Đáp Thực vật Rừng Khu vực Đà Nẵng” (FloraQA) đã hoàn thành các mục tiêu và chức năng chính đã đề ra. Tuy nhiên, như bất kỳ công trình nghiên cứu nào, vẫn còn những khía cạnh cần được hoàn thiện và cải tiến hơn nữa trong tương lai.

### Thành tựu và Đóng góp chính

- **Về mặt lý thuyết:**

- Đã tiến hành nghiên cứu sâu rộng các tài liệu liên quan, kỹ thuật và các hệ thống nhận diện thực vật hiện có, qua đó có được sự hiểu biết toàn diện về các phương pháp tiên tiến nhất.
- Nâng vững kiến thức lý thuyết chuyên sâu về phương pháp xây dựng cơ sở tri thức chuyên biệt, chiến lược huấn luyện và xây dựng mô hình (đặc biệt là Dual-Stream Fusion, ProxyNCA Loss), các thuật toán liên quan (OOD detection) và kỹ thuật triển khai hệ thống trong thực tế.
- Chủ động khám phá các công nghệ mới và xác định các xu hướng nghiên cứu chính trong lĩnh vực nhận diện thực vật đa phương thức và hệ thống hỏi-đáp tăng cường truy xuất (RAG), góp phần vào sự hiểu biết sâu rộng hơn về lĩnh vực này.

- **Về mặt thực tiễn:**

- **Kiến trúc hệ thống mới:** Thiết kế và triển khai thành công một kiến trúc hệ thống hỏi - đáp đa phương thức (ảnh và văn bản) toàn diện, đạt được độ chính xác cao và có khả năng mở rộng cho các ứng dụng thực tế.
- **Cải tiến trong phân loại:** Phân tích và tích hợp kiến trúc Dual-Stream Fusion, kết hợp thế mạnh của ConvNeXt V2 và Vision Transformer, giúp cải thiện đáng kể khả năng phân loại các loài thực vật có hình thái phức tạp và biến thiên, vượt trội so với các mô hình đơn lẻ có cùng quy mô.

- **Phát hiện mẫu ngoài miền hiệu quả:** Xây dựng và áp dụng thuật toán phát hiện mẫu ngoài phân phôi (OOD) dựa trên trung bình khoảng cách theo lớp, tăng cường đáng kể độ tin cậy và an toàn của hệ thống khi đối mặt với dữ liệu không xác định trong thực tế.
- **Xây dựng cơ sở tri thức chuyên biệt:** Xây dựng một cơ sở tri thức độc đáo và chi tiết về thực vật rừng khu vực Đà Nẵng, đặc biệt là các thông tin về y học cổ truyền được trích xuất và chuẩn hóa từ các nguồn tài liệu khoa học uy tín, làm giàu nguồn dữ liệu cho hệ thống.
- **Hiệu suất cao:** Chứng minh hiệu quả của hệ thống FloraQA thông qua đánh giá nghiêm ngặt, đạt được kết quả ấn tượng (Accuracy@5 đạt 87.05% cho phân loại, AUROC đạt 0.9359 cho lọc OOD), cho thấy tiềm năng ứng dụng thực tiễn to lớn.
- **Giao diện thân thiện:** Phát triển một giao diện web trực quan và dễ sử dụng, tạo điều kiện cho người dùng phổ thông, sinh viên và nhà nghiên cứu dễ dàng tiếp cận và khai thác hệ thống FloraQA.

### **Hạn chế**

- **Hạn chế về bộ dữ liệu:** Dù có số lượng lớn, bộ dữ liệu vẫn còn mất cân bằng và chủ yếu tập trung vào ảnh lá cây khỏe mạnh, thiếu sự đa dạng về các bộ phận khác (hoa, quả), các giai đoạn phát triển và các trường hợp cây bị bệnh. Ngoài ra, sự chênh lệch về số lượng ảnh giữa các loài cũng gây nên những khó khăn trong quá trình huấn luyện mô hình, khi các mô hình sẽ thường thiên vị các loài có số lượng ảnh lớn trong tập huấn luyện.
- **Phạm vi cơ sở tri thức:** Cơ sở tri thức hiện tại chủ yếu tập trung vào các loài ở Đà Nẵng và ứng dụng y học, cần được mở rộng để bao quát các khía cạnh khác như sinh thái học, tình trạng bảo tồn và các loài ở khu vực lân cận.
- **Phụ thuộc vào dịch vụ bên ngoài:** Việc sử dụng API Gemini cho chức năng sinh văn bản tạo ra sự phụ thuộc, có thể phát sinh chi phí và bị giới hạn bởi chính sách của nhà cung cấp.

- **Khả năng xử lý câu hỏi phức tạp:** Hệ thống định tuyến câu hỏi dựa trên quy tắc có thể chưa xử lý tốt các câu hỏi đa ý, mơ hồ, yêu cầu tính liên kết với câu hỏi trước hoặc đòi hỏi suy luận sâu.

### **Hướng phát triển trong tương lai**

- **Mở rộng và đa dạng hóa bộ dữ liệu:** Hợp tác với các nhà thực vật học và cộng đồng để thu thập một bộ dữ liệu lớn hơn, đa dạng hơn và cân bằng hơn, bao gồm nhiều bộ phận, điều kiện và giai đoạn phát triển của cây để tăng khả năng khái quát hóa của mô hình.
- **Cải thiện khả năng xử lý ngôn ngữ:** Nâng cấp khả năng định tuyến câu hỏi với các công nghệ tiên tiến như hệ thống đa tác nhân (Multi-agent system) thay vì rule-based, tăng tính linh hoạt trong khả năng hiểu câu hỏi.
- **Nâng cao cơ sở tri thức:** Tích hợp thêm nhiều nguồn dữ liệu khoa học để mở rộng cơ sở tri thức, bổ sung thông tin về sinh thái, phân bố địa lý, và cập nhật các thay đổi về phân loại học.
- **Tối ưu hóa mô hình và hệ thống:** Nghiên cứu các chiến lược fine-tune LLM cục bộ (khi tài nguyên cho phép) để giảm sự phụ thuộc và tăng khả năng tùy chỉnh. Cải thiện thuật toán OOD để xử lý tốt hơn các loài có quan hệ họ hàng gần.
- **Mở rộng ứng dụng:** Phát triển ứng dụng di động dựa trên hệ thống đã xây dựng để hỗ trợ nhận diện thực vật tại thực địa. Cung cấp API để các nhà nghiên cứu hoặc các ứng dụng khác có thể tích hợp và sử dụng.
- **Tích hợp cơ chế phản hồi từ cộng đồng:** Xây dựng tính năng cho phép người dùng đóng góp dữ liệu, báo cáo nhận diện sai và xác thực thông tin, tạo ra một vòng lặp cải tiến liên tục cho hệ thống.

## TÀI LIỆU THAM KHẢO

- [1] Gamfeldt, L., Snäll, T., Bagchi, R., Jonsson, M., Gustafsson, L., Kjellander, P., ... & Bengtsson, J., «Higher levels of multiple ecosystem services are found in forests with more tree species,» *Nature Communications*, vol. 4, n° %11, p. 1340, 2013.
- [2] Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., ... & Hayes, D., «A large and persistent carbon sink in the world's forests,» *Science*, vol. 333, n° %16045, pp. 988-993, 2011.
- [3] «Conservation work in Vietnam,» *Fauna & Flora International*, 2015. [En ligne]. Available: <https://www.fauna-flora.org/countries/vietnam/>.
- [4] «The Importance of Forests,» *WWF*, 2024. [En ligne]. Available: [https://wwf.panda.org/discover/our\\_focus/forests\\_practice/importance\\_forests/](https://wwf.panda.org/discover/our_focus/forests_practice/importance_forests/).
- [5] Fikret Berkes, *Sacred Ecology*, Routledge, 2018.
- [6] Wäldchen, J., & Mäder, P., «Machine learning for image based species identification,» *Methods in Ecology and Evolution*, vol. 9, n° %111, pp. 2216-2225, 2018.
- [7] K. M. Parsley, «Plant awareness disparity: A case for renaming plant blindness,» *Plants, People, Planet*, vol. 2, n° %16, pp. 598-601, 2020.
- [8] Vandebroek, I., Calewaert, J. B., De jonckheere, S., Sanca, S., Semo, L., Van Damme, P., ... & Reyes-García, V., «Use of medicinal plants and pharmaceuticals by indigenous communities in the Bolivian Andes and Amazon,» *Bulletin of the World Health Organization*, vol. 82, n° %14, pp. 243-250, 2004.
- [9] D. N. P. Committee, «Master Plan on Biodiversity Conservation in 2030.,» 2020. [En ligne].
- [10] T. Mucci, «What is Question Answering?,» IBM, 2025. [En ligne]. Available: <https://www.ibm.com/think/topics/question-answering>.

- [11] J. O. Schneppat, «Question Answering (QA): Challenges and Approaches,» Schneppat Technology Solutions, 2024. [En ligne]. Available: [https://schneppat.com/question-answering\\_qa.html](https://schneppat.com/question-answering_qa.html).
- [12] N. Klingler, «Understanding Visual Question Answering (VQA) in 2025,» viso.ai, 2025. [En ligne]. Available: <https://viso.ai/deep-learning/understanding-visual-question-answering-vqa>.
- [13] Zhou, J., Li, J., Wang, C., Wu, H., Zhao, C., & Teng, G., «Crop disease identification and interpretation method based on multimodal deep learning,» Computers and Electronics in Agriculture, vol. 189, 2021.
- [14] Kolluri, J., Dash, S. K., & Das, R., «Plant Disease Identification Based on Multimodal Learning,» International Journal of Intelligent Systems and Applications in Engineering, vol. 12, n° % 14, pp. 485-492, 2024.
- [15] Yang, X., Gao, J., Xue, W., & Alexandersson, E., «PLLaMa: An Open-source Large Language Model for Plant Science,» arXiv preprint arXiv:2401.01600, 2024.
- [16] Wäldchen, J., Rzanny, M., Seeland, M., & Mäder, P., «Automated plant species identification—Trends and future directions,» PLOS Computational Biology, vol. 14, n° % 14, 2018.
- [17] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K. W., Zhu, S. C., ... & Kalyan, A., «Learn to explain: Multimodal reasoning via thought chains for science question answering,» chez Advances in Neural Information Processing Systems, 2022.
- [18] Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T. W., McNeill, J., Monro, A. M., Prado, J., Price, M. J. & Smith, G. F., International Code of Nomenclature for Algae, Fungi, and Plants (Shenzhen Code), Koeltz Botanical Books, 2018.
- [19] «Nomenclature: Understanding Plant Names,» Pennsylvania State University Extension, 2024. [En ligne]. Available: <https://extension.psu.edu/nomenclature>.

- [20] Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., ... & Boujema, N., «Interactive plant identification based on social image data,» Ecological Informatics, vol. 23, pp. 22-34, 2014.
- [21] Hart, A. G., Bosley, A., Hooper, T., Mitchell, R., & Goodenough, A. E., «Assessing the accuracy of free automated plant identification applications,» People and Nature, vol. 5, n° %14, pp. 1191-1197, 2023.
- [22] Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S., «The iNaturalist species classification and detection dataset,» chez Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
- [23] «New Computer Vision Model (v2.12) with 1,983 new taxa,» iNaturalist, 2024. [En ligne]. Available: <https://www.inaturalist.org/blog/91824-new-computer-vision-model-v2-12>.
- [24] August, T., Harvey, M., Lightfoot, P., Kilbey, D., Papadopoulos, T., & Jepson, P., «AI naturalists might hold the key to unlocking biodiversity data in social media imagery,» Patterns, vol. 1, n° %17, 2020.
- [25] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D., «Retrieval-augmented generation for knowledge-intensive nlp tasks,» chez Advances in Neural Information Processing Systems, 2020.
- [26] Gillespie, L. E., Rodríguez-González, A., Boer, M. M., Fairman, T. A., Hui, C., Rumpf, S. B., ... & Expósito-Alonso, M., «Deep learning models map rapid plant species changes from citizen science and remote sensing data,» Proceedings of the National Academy of Sciences, vol. 121, n° %137, 2024.
- [27] Loarie, S. R., Jetz, W., & Mace, G. M., «The computer vision model,» iNaturalist Computer Vision Explorations, 2017. [En ligne]. Available: [https://www.inaturalist.org/pages/computer\\_vision\\_demo](https://www.inaturalist.org/pages/computer_vision_demo).
- [28] Bartlett, P., White, K. M., Johnstone, D., Borgen, T., Heilmann-Clausen, J., & Frøslev, T. G., «Using computer vision to automate the identification of Hebeloma

- mushroom species,» Methods in Ecology and Evolution, vol. 14, n° %11, pp. 217-228, 2023.
- [29] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., & Xie, S., «ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders,» chez Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
- [30] He, K., Zhang, X., Ren, S., & Sun, J., «Deep residual learning for image recognition,» chez Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
- [31] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N., «An image is worth 16x16 words: Transformers for image recognition at scale,» chez International Conference on Learning Representations, 2021.
- [32] J. Zhang, «Weed Recognition Method based on Hybrid CNN-Transformer Model,» *Frontiers in Computing and Intelligent Systems*, vol. 4, n° %12, 2023
- [33] Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., & Singh, S., «No fuss distance metric learning using proxies,» chez Proceedings of the IEEE International Conference on Computer Vision, 2017.
- [34] Kim, S. E., Kim, G. H., & Heo, J. P., «ProxyNCA++: Revisiting and revitalizing proxy neighborhood component analysis,» chez Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020.
- [35] Johnson, J., Douze, M., & Jégou, H., «Billion-scale similarity search with GPUs,» *IEEE Transactions on Big Data*, vol. 7, n° %13, pp. 535-547, 2019.
- [36] «Distributional semantics,» Wikipedia, [En ligne]. Available: [https://en.wikipedia.org/wiki/Distributional\\_semantics](https://en.wikipedia.org/wiki/Distributional_semantics).
- [37] Lenci, A., & Sahlgren, M., Distributional Semantics (Studies in Natural Language Processing), Cambridge University Press, 2023.

- [38] Amigó, E., Ariza-Casabona, A., Fresno, V., & Martí, M. A., «Information Theory-based Compositional Distributional Semantics,» Computational Linguistics, vol. 48, n° %14, p. 907–948, 2022.
- [39] «Vector Embeddings Explained,» Weaviate, [En ligne]. Available: <https://weaviate.io/blog/vector-embeddings-explained>.
- [40] K. You, «Semantics at an Angle: When Cosine Similarity Works Until It Doesn't,» arXiv preprint arXiv:2504.16318, 2024.
- [41] «What Is Retrieval-Augmented Generation aka RAG,» NVIDIA, 2025. [En ligne]. Available: <https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/>.
- [42] «Retrieval-augmented generation,» Wikipedia, [En ligne]. Available: [https://en.wikipedia.org/wiki/Retrieval-augmented\\_generation](https://en.wikipedia.org/wiki/Retrieval-augmented_generation).
- [43] «Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement,» Nature Human Behaviour, 2024. [En ligne]. Available: <https://www.nature.com/articles/s41562-024-02089-y>.
- [44] «Improving Equity and Access to Non-English Large Language Models,» Stanford HAI, [En ligne]. Available: <https://hai.stanford.edu/news/improving-equity-and-access-non-english-large-language-models>.
- [45] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I., «Attention is all you need,» chez Advances in Neural Information Processing Systems, 2017.
- [46] R. M. Schmidt, «Recurrent Neural Networks (RNNs): A gentle introduction and overview,» arXiv preprint arXiv:1912.05911, 2019.
- [47] Touvron, H., Cord, M., Douze, M., Jégou, H., & Sablayrolles, A., «Training data-efficient image transformers & distillation through attention,» chez International Conference on Machine Learning, 2021.
- [48] He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R., «Masked autoencoders are scalable vision learners,» chez Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition, 2022.

- [49] «Massive Text Embedding Benchmark (MTEB) Leaderboard,» Hugging Face, [En ligne]. Available: <https://huggingface.co/spaces/mteb/leaderboard>.
- [50] Chen, Q., Wu, X., Wu, Y., Liu, H., Lv, J., & Lin, L., «BGE-M3: A Massive Multilingual, Multigranularity, and Multimodal Embedding Model,» arXiv preprint arXiv:2402.03216, 2024.
- [51] Đ. T. Lợi, *Những cây thuốc và vị thuốc Việt Nam*, Nhà xuất bản Khoa học và Kỹ thuật, 2004.
- [52] G. Gemini Team, «Gemini: A family of highly capable multimodal models,» arXiv preprint arXiv:2312.11805, 2023.
- [53] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D., «Grad-CAM: Visual explanations from deep networks via gradient-based localization,» chez Proceedings of the IEEE International Conference on Computer Vision, 2017.
- [54] Cormack, G. V., Clarke, C. L. A., & Buettcher, S., «Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods,» chez Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009.
- [55] K. Sparck Jones, «A statistical interpretation of term specificity and its application in retrieval,» Journal of Documentation, vol. 28, n° %11, pp. 11-21, 1972.
- [56] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M., «Okapi at TREC-3,» chez The Third Text REtrieval Conference, 1995.
- [57] Gao, L., Dai, Z., & Callan, J., «Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval,» chez Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.
- [58] Jeff Johnson, Matthijs Douze, Hervé Jégou, «Billion-scale similarity search with GPUs,» chez IEEE Transactions on Big Data, 2017.

Hệ thống hỏi đáp thực vật rừng khu vực Đà Nẵng